ENHANCING EXPRESSIVITY OF DOCUMENT-CENTERED COLLABORATION WITH MULTIMODAL ANNOTATIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Dongwook Yoon

August 2017

ENHANCING EXPRESSIVITY OF DOCUMENT-CENTERED COLLABORATION WITH MULTIMODAL ANNOTATIONS

Dongwook Yoon, Ph.D.

Cornell University 2017

As knowledge work moves online, digital documents have become a staple of human collaboration. To communicate beyond the constraints of time and space, remote and asynchronous collaborators create digital annotations over documents, substituting face-to-face meetings with online conversations. However, existing document annotation interfaces depend primarily on text commenting, which is not as expressive or nuanced as in-person communication where interlocutors can speak and gesture over physical documents. To expand the communicative capacity of digital documents, we need to enrich annotation interfaces with face-to-face-like multimodal expressions (e.g., talking and pointing over texts). This thesis makes three major contributions toward multimodal annotation interfaces for enriching collaboration around digital documents.

The first contribution is a set of design requirements for multimodal annotations drawn from our user studies and explorative literature surveys. We found that the major challenges were to support lightweight access to recorded voice, to control visual occlusions of graphically rich audio interfaces, and to reduce speech anxiety in voice comment production. Second, to address these challenges, we present

RichReview, a novel multimodal annotation system. RichReview is designed to capture natural communicative expressions in face-to-face document descriptions as the combination of multimodal user inputs (e.g., speech, pen-writing, and deictic pen-hovering). To balance the consumption and production of speech comments, the system employs (1) cross-modal indexing interfaces for faster audio navigation, (2) fluid document-annotation layout for reduced visual clutter, and (3) voice synthesis-based speech editing for reduced speech anxiety. The third contribution is a series of evaluations that examines the effectiveness of our design solutions. Results of our lab studies show that RichReview can successfully address the above mentioned interface problems of multimodal annotations. A subsequent series of field deployment studies test the real-world efficacy of RichReview by deploying the system for document-centered conversation activities in classrooms, such as instructor feedback for student assignments and peer discussions about course material. The results suggest that using rich annotation helps students better understand the instructor's comments, and makes them feel more valued as a person. From the results of the peer-discussion study, we learned that retaining the richness of original speech is the key to the success of speech commenting. What follows is the discussion on the benefits, challenges, and future of multimodal annotation interfaces, and technical innovations required to realize the vision.

BIOGRAPHICAL SKETCH

Yoon's research lies at the intersection of human-computer interaction, computer-supported cooperative work, computer-mediated communication, and educational technology. He builds interactive systems powered by expressive multimodal interactions. His work frequently appears at the top-tier ACM venues including UIST, CSCW, and CHI. His multimodal document commenting system, RichReview, has been deployed to six different classes at Cornell, including a massive open online course where it was successfully used by students around the world as a tool for online peer discussion assignment. He has worked at Microsoft Research, edX, and KIST. He earned his B.S. in electrical engineering and M.S. in computer graphics from Seoul National University. His Ph.D. study was supported by the Kwanjeong Scholarship.

To Mihae

ACKNOWLEDGMENTS

The successful completion of this dissertation was made possible by the support of my advisors, mentors, friends, and family. I made it through this hardest—but most rewarding—time of my life thanks to their encouragement, advice, and never-ending love.

I would like to extend my huge gratitude to my advisor, François Guimbretière. His high standards and resolution motivated me to tackle ambitious research problems. When hard times hit, he offered his unwavering belief and unsparing support for me until I made it through. His advice helped me achieve a research goal that I am genuinely proud of, and enabled me to grow as a researcher and as a person.

I would also like to thank the members of my committee who brought indispensable expertise and experience into my work. Susan Fussell guided me back to the most fundamental and important matters. Erik Andersen has always been a valuable source of fresh perspectives on my research.

My internship mentors in Microsoft Research and edX deserve recognition as well. Nicholas Chen strengthened my research abilities from the inception of my thesis and often felt like a co-advisor of my Ph.D. study. Abigail Sellen embodied what it means to be a true mentor. Ken Hinckley helped widen my perspective on my academic career. Piotr Mitros inspired me to envision the future of educational technology.

My research has given me the chance to engage with a number of awesome collaborators. Corinna Löckenhoff was the first to cast light on the real-world applicability of my work. Steven Jackson provided incredible high-level insights for my deployment studies, and his counsel helped me through the hardest time of my Ph.D. study. Michel Pahud demonstrated professionalism and selfless devotion. Venkatesh Sivaraman and Ian Arawjo helped me grow as a mentor by patiently following my lead.

Many people supported my running of user studies and fieldwork. People from Cornell Academic Technology took the risk of deploying my systems to classes across the campus and CornellX. Instructors, including Maria Terrell and Maria Wolfe, students, and participants in my fieldwork and lab studies helped me find system defects and provided insightful reports for better design.

I would like to gratefully acknowledge the generous financial support from many institutions. The Kwanjeong Educational Foundation has covered my living expenses throughout my Ph.D. Execution of my study was possible in part by gifts from Microsoft and FXPAL.

I feel very fortunate to have the great friends at Cornell Information Science and Korea Graduate Student Association. I humbly realize that it would be impossible to mention by name every other individual who made beautiful memories of Ithaca together, but they will know. Thank you for sharing all the wine and beer!

Thanks to my family, for being there always. Mom, Dad, and Jungwon hold belief in me, and supported me every step of the way. My love and my best friend, Mihae, stood by me through the hard times, good times, and so many other times. I am so lucky to have you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

Face-to-face interaction, as compared to meeting in a digital setting, offers unmatched expressivity for conveying complex ideas and nuanced emotions (e.g., emotions embedded in voice inflection or the unspoken meaning of a pointed finger). To enhance expressivity of digital annotation tools, several studies have leveraged multimodal user inputs, i.e., interactions through multiple communication channels, such as speech (Neuwirth et al., 1994), writing (Anderson et al., 2006), and gesture (Tsang et al., 2002), but the main challenge is that people using the systems have difficulty creating, managing, and sharing the resulting multimedia comments (e.g., editing a recorded voice comment is not as easy as editing text).

To design, implement, demonstrate, and evaluate solutions for such problems, we built a multimodal annotation system called RichReview. The purpose of our systems work is to evaluate the following thesis:

> *New multimodal interaction techniques will make document commenting a viable alternative to face-to-face conversation in educational settings.*

This dissertation makes three contributions on the way toward evaluating this thesis. We *invented new techniques for multimodal commenting* that break through limitations of existing systems. The subsequent lab studies and fieldwork validated our new solutions by *examining the relevant human factors*, such as speech anxiety and perceived efficacy. Then we conducted a series of field deployments to test the *real-world efficacy* of our solutions. This new knowledge will help us reimagine what

digital tools can offer to enable collaborators to work beyond the barriers of space and time.

Promises and pitfalls of digital tools for document-centered collaboration

The expressivity and richness of an in-person meeting draws primarily on its multiple communication channels (Clark, 1996). An exemplary use case of the collocated multimodal interactions is document-centered conversation. When talking about a shared document, people *speak* about the texts while *gesturing* over the part of the page, making a visual reference (Bickmore et al., 2008). They also *write* markups on texts to help the collaborators keep track of the points they made (Sellen & Harper, 2003). Yet, despite the unmatched expressive power of face-to-face meetings—the combination of voice, gesture, and inking modalities—physical in-person meetings have the crucial constraints of time and space due to its synchronous and collocated nature.

When a face-to-face meeting is undesirable, many digital annotation tools have been built to connect asynchronous and distant document workers beyond the spatiotemporal barriers. Digital annotation tools have several advantages when compared to face-to-face meetings. First, collaborators can work together without physically being together or attending the meeting at the same time, which offers unmatched flexibility in the work process. Second, the asynchronous and remote nature reduces resources and costs required for collaboration. And finally, the collaborators can work at their own pace, because worker activities in the asynchronous systems are archived for private revision and evaluation in the future.

Therefore, the existing tools rely heavily on textual communication which lacks the natural communication capability that everyone has. They are missing the benefits of multimodal interactions, since most desktop or laptop software is built based on keyboard and mouse inputs. To go beyond the graphical user interface paradigm, several previous systems have incorporated voice, inking, or gesture interactions (Levine & Ehrlich, 1991; Neuwirth et al., 1994; Tsang et al., 2002). However, rich and expressive annotation systems have not merged into our everyday digital lives yet.

The core challenge is that the richness of the multimodal content comes at the excessive expense of user workloads. For instance, displaying the visual interfaces to the multimedia content (e.g., written annotations, navigation controls, and editing menus) can impose a cognitive burden on the user by cluttering screen real estate. In the course of exploring solutions to this problem, we start by focusing on *inking*, which is the most basic and popular mode of multimodal interaction as it appears in a large body of human–computer interaction (HCI) studies (Anderson, 2004; Hinckley et al., 2007; Marshall, 1998; Schilit et al., 1998).

Creating space in a digital document for presenting rich visual content

Writing on a page is a lightweight strategy to create fluid and rich visual expressions on a document (Marshall, 1997; Price et al., 1998). For inking, *whitespace* is a crucial real estate resource that offers ample room for writing, and contextualizes the written content with adjacent body texts (Marshall, 1998). However, whitespace is also a scarce resource in static documents where people often run out of space for taking notes.

Our solution to this real estate problem is TextTearing, an interaction technique for creating new whitespace between the text lines in dynamic document layouts (Yoon et al., 2013). To best leverage the interactive capacity of tablet computers, we designed and built four different gesture options powered by varying combinations of pen and touch inputs. The lab study compared these gesture options against the baseline of a vanilla pen writing interface. The results showed that the pen-only pigtail gesture called PenTearing is easier to learn, faster, and more lightweight than the alternatives. Hence, in the next step of our study, we decided to retain the fluid document layout feature along with the PenTearing interaction to accommodate the interfaces for displaying rich multimodal content (e.g., audio waveforms and whitespace for inking).

Promoting fast and direct access to multimodal comments

The most imminent and apparent interface problem of multimodal commenting systems is heavy workload of consuming rich content (Grudin, 1988). The bare necessity for the recipient of a multimodal comment is a set of features for browsing and skimming the rich streams of data that include voice recording and time-stamped gestures. Listening to speech data is especially tedious and time-consuming because the listener should listen to the recording from beginning to the end to know what is where. What is lacking in conventional audio interfaces (e.g., a slider bar in addition to play/pause/stop buttons) is visual cues for navigation. Unlike speech, text is easy to consume—read—because written words by themselves serve as visual cues for semantic navigation.

To solve the consumption problem, we designed the voice interfaces of RichReview to employ rich visual presentations, such as audio waveform or auto-transcriptions (Yoon et al., 2014). Waveform provides a direct visual proxy for visual navigation and skimming of voice, as the user can look and spot where pauses start and end. For a semantic browsing, we can turn waveforms into auto-caption words if automatic speech recognition is available.

However, there's a design dilemma in the visual interfaces for speech annotation. A comment on a page can be given context by anchoring it to a given part of the text (e.g., "This paragraph needs revision."). Putting the comment adjacent to anchor text enhances the context, but presenting the visual interface can occlude the body text because it demands a larger space than a slider bar. Several commodity applications anchor small voice icons that expand to an on-demand visual interface, but clicking the small buttons to toggle the interfaces slows users down. The gist of this problem is that there's not enough room to display the rich representation of speech comments near the corresponding body text.

We broke through this real estate dilemma by employing TextTearing, the fluid document layout technique. Creating a RichReview comment reflows the lines of the column as if the waveform or transcription interface is a part of the body text. *Fast and direct access* to any part of the voice is guaranteed, since every line of waveform and transcription is always open, just like the body text. Accordingly, with a lightweight tapping interaction, the user can *index* any part of the visual proxy to listen to the corresponding part of the voice recording. In addition, collaborators can maintain conversational threads by juxtaposing multiple in-line comments. A series of

follow-up studies testifies to the efficacy of our design for easy consumption of multimodal comments, especially in an educational setting. In a formative evaluation, novice users reported that they could use a versatile mixture of speech, gesture, and inking to express complex ideas, and that the indexing features helped them quickly navigate through recorded multimodal comments.

Testing real-world efficacy of RichReview in classroom settings

To examine the real-world efficacy of our new multimodal commenting system, we conducted a series of field studies by deploying RichReview to several classes on the Cornell campus. We first targeted a way to enhance the instructor feedback process, because a teacher's comments to students in writing is an important and widespread document-centered collaboration activity in classrooms. Our first deployment was for a term paper feedback process in a small social science class. The students found RichReview comments easier to understand than longhand writing on paper thanks to delivered emotion and nuances of speech. Moreover, they even preferred RichReview over office hours, because replaying asynchronous comments allowed them to examine the instructor's feedback at their pace.

In the second deployment, we extended the use case of RichReview to assignments and prelim feedback in a large math class. The reports from the students confirmed the core benefits of the system observed from the previous study. Also, the students felt that they were valued as a person because spoken comments felt more personal and careful than pen writing. In addition, they could perceive the full benefits of

RichReview feedback only when the instructors exploited the full capability of the system by recording a lengthy comment.

Extending the task context to online peer discussion

Inspired by the success from the instructor feedback use cases, our next study targeted RichReview for peer discussions, another wide-spread document-centered collaboration in classrooms. This transition poses new opportunities and challenges, because the dynamics of peer discussion differ from that of instructor feedback. Instructor feedback is a unidirectional pedagogy that doesn't scale well to the massive open online course (MOOC) size classes due to the very high student-to-instructor ratio. In contrast, discussants can take the homogeneous role which makes it applicable to a very large course by structuring students into small discussion groups. However, the challenge is that now everybody speaks. To support this bidirectional communication model, we would build a new web-based collaboration system, RichReview.net, where students can create discussion threads of multimodal comments over shared documents using their own laptops, since providing tablets for all students would be too costly.

For the peer discussion study, we deployed RichReview.net to a small social science course for a weekly online discussion. From the study, we found new challenges in *production* of speech comments. Students felt speech comments took more effort to create than textual ones. Qualitative investigation revealed three major problems in speech commenting: (1) recorded speech was harder to edit than text, (2) students felt speech anxiety due to potential disfluencies and lack of anonymity, and

(3) they didn't want to hear their own voice. To tackle these new challenges, we redesigned and rebuilt the speech commenting interface taking a fresh approach of speech re-synthesis.

TypeTalker: an interface for reducing workloads and anxiety of speech commenting

On the production side, we focused on reducing workloads of voice editing and speech anxiety of live recording. Our speech production interface employed the combination of automatic speech recognition and speech synthesis techniques to tackle these problems. Using caption words of speech as textual proxies enabled word-level voice editing through keyboard interaction, which was much easier and faster than traditional timeline-based waveform editing. Surrogating the user's voice with a synthesized voice simplified the production process, because inserting a new segment of speech in the middle of an existing voice stream can be done though lightweight keyboard input without re-recording. Moreover, the synthesized voice can reduce a speaker's anxiety or self-consciousness as it can disguise the speaker's identity. To test if our speech synthesis-based solution can reduce user workloads and anxiety, we conducted a production-side evaluation where participants create speech comments. From the results, we observed that the surrogated voice not only resolved a speaker's self-consciousness, but also reduces user workload in revision practices, such as insertion and small edits.

Deploying RichReview for online peer discussion in a MOOC

The subsequent step after exploring the interface solution for the speech production problems is to test it in the wild. For evaluating the real-world efficacy of

our synthesis-based approach, we updated a RichReview discussion system with the TypeTalker feature with scalable, secure, and accessible infrastructure for a large-scale deployment. From the results of the study, we could observe both the promises and challenges of rich commenting for peer discussion. When a student was in a setting where she can use the speech commenting, it motivated her to participate in the discussion more actively. Yet, we also learned that there are a couple of remaining technical challenges for wide acceptance of speech, as users often face issues of low speech recognition rate and the low-quality audio recording on the way to using speech.

Promises and challenges of multimodal annotation, and implications for future studies

Chapter 8 builds on the findings from our work and reflects on several discussion points along with opportunities for future work. We first present implications of the newly found benefits and challenges for the design of the rich commenting system. The discussion suggests that the beneficial features of our tool (e.g., asynchronicity) match specific needs emerging from the ecological settings (e.g., self-paced learning in an educational setting) under the condition that proper design of the system successfully addressed the core challenges of using rich media (e.g., access, editing, anxiety, editing effort). What follows is a meta-analysis of existing speech commenting systems, including ours, that sheds light on the ideation of the optimal speech commenting system that can satisfy the user needs on the both sides of communication: commentator and listener. Finally, we discuss the remaining technical challenges in bringing multimodal commenting into the everyday digital lives of

people. In recollection of the findings from our deployment studies, we specifically highlight the needs for accurate speech recognition and quality audio recording.

In the field of educational technology, multimodal annotation is a relatively unexplored topic. This opens opportunities for application of rich commenting for classroom tasks beyond the active reading that we have explored throughout this dissertation. Therefore, we discuss how the implications from the past RichReview studies can be transferred to other use cases on campus, such as collecting course evaluations or providing emotional support. Furthermore, although we built and evaluated RichReview as a classroom tool, there are generalizable findings about multimodal annotation that are worth exploring in other contexts. We thus envision the broader impacts of multimodal annotation to new settings, including code review and virtual reality applications.

2 Related works

In an attempt to put this dissertation in the context of previous work in the field of HCI, this chapter reflects on the literature in the subfields related to the topic of this thesis including interaction design, computer-mediated communication, and computer-supported cooperative work. We chronicle the evolutions of inquiries and approaches toward multimodal commenting systems by revisiting the core interaction modes that have been commonly used to improve digital document annotation. We start with digital inking, the most basic way to markup text. Then we discuss speech, a high-throughput communication channel. The next section focuses on deictic gesture, a visual mode for connecting speech to text. And lastly, we go over the applications of rich commenting systems in educational settings. This review will help the reader understand the rest of this dissertation.

2.1 Inking[1]

Freeform ink annotations are pervasive and used extensively for document work because they are fast to create, can be interleaved with the reading process (Sellen & Harper, 2003), and are highly flexible in the information they represent (Marshall & Brush, 2004). As a result, several annotation systems in the literature have employed ink as a primary modality. The collaborative editor MATE (Hardock et al., 1993) supported the use of ink for low-level editing commands as well as served as a general

[1] The text of this section was derived from previous publications (Yoon et al., 2013, 2014).

medium for communication. Similarly, the XLibris reading device, which supported pen input, was used to explore various use scenarios of collaborative ink annotations (Marshall et al., 1999). Having margins for whitespace is the crucial affordance of document pages that helps contextualize written annotations in relation to nearby text (Marshall, 1998). However, pages have a limited amount of space in the margins for writing (Pearson et al., 2009). To bypass the physical constraints of paper documents, researchers in HCI have explored interface solutions as follows.

Chang et al. (1998) introduced the notion of fluid documents, which dynamically adjust their layout in order to display secondary information. Design of our thesis system follows this strategy but apply it in the context of annotation creation. Zeleznik et al. (2010) created a bimanual gesture for inserting and removing whitespace within a digital canvas to solve similar problems of limited writing space encountered when working through math problems. In their system, this feature is activated when a touch following a pen stroke manipulates a feed forward widget. We used a simplified version of this approach in the initial design of our system. Our work provides performance data about different variations of this interaction technique. LiquidText (Tashman & Edwards, 2011) used various multi-touch gestures to collapse, rearrange, highlight and extract portions of a document and GatherReader (Hinckley et al., 2012) explored how these tasks could be further enhanced using pen + touch interactions.

Although the interactions in the previous systems bear a resemblance to our solution technique called TextTearing, they serve fundamentally different purposes. In these other systems, the interactions assist in juxtaposing or gathering different parts

of a document, whereas the interactions in TextTearing are designed to help the user make in situ ink annotations.

2.2 Speech

Speech has been employed as the central element of most multimodal systems for its many beneficial characteristics. In the literature of organizational and business communication, collaboration over voice media, compared with text, is known to yield faster decision making (Williams, 1977) and reduced equivocality (Daft & Lengel, 1986). Standing on these promises, HCI researchers studied how recorded speech messages can enhance digital communication by possessing high throughput (Grudin, 1988), delivering nuance and emotions (Chalfonte et al., 1991; Yaneske & Oates, 2010), clarifying the speaker's intention (Hew & Cheung, 2013), supporting a positive perception of the speaker (Neuwirth et al., 1994; Oomen-Early et al., 2008), and addressing higher-level concerns such as semantic and structural aspects more effectively than text-based comments (Chalfonte et al., 1991; Kraut et al., 1992; Neuwirth et al., 1994). Despite these advantages, spoken annotation is yet to be widely used in our everyday lives. Previous studies identified the two major challenges of speech: access and production.

2.2.1 ACCESSING SPEECH

Jonathan Grudin (1988) pioneered studies on issues surrounding voice application in a cooperative setting. His study was the first to raise awareness about the speech consumption problem: browsing often takes longer than speaking. He suggested that the linear nature of recorded audio is the major reason why it does not readily offer

quick skimming and browsing of its contents. Multimodal annotation systems that use voice as a central mode of interaction inevitably inherit this problem.

HCI researchers have explored three types of solutions to enhance access to speech by structuring the linear data into types of higher level constructs. The first approach is acoustic structuring that extracts navigational cues from features of the audio signal, such as absence of speech and change of inflection (Arons, 1993; Hindus & Schmandt, 1992; Schmandt, 1981). The basic method of acoustic structuring is to offer access to voice through waveform interface. However, the utility of using waveforms to navigate annotations largely remained unnoticed in the field. In this dissertation, we conducted a lab experiment (Yoon et al., 2014) as well as a deployment study (Yoon et al., 2016) to reveal the benefits of navigating multimodal annotations using audio waveform.

The second category of solution is semantic structuring that uses captions from auto transcription as a visual proxy for speech (Monserrat et al., 2013; Whittaker et al., 2002). Although skimming the caption words enables fast and parallel access to speech, captions from automatic speech recognition (ASR) have transcription errors by nature. On one hand, words from the error-laden transcript have been shown to act as semantic cues for extracting key points or as navigational cues for browsing (Whittaker et al., 2002), but on the other hand, recognition errors hurt listener comprehension (Stark et al., 2000), especially for non-native speakers (Pan et al., 2009). When the recognition rate is very low, listener comprehension might be as bad as (Vemuri et al., 2004) or even worse than (Munteanu et al., 2006) no transcription (e.g., reading homophone transcription errors can induce confusion by misleading

listening comprehension). As a solution, Vemuri et al. (2004) suggested confidence shading that diminishes the opacity of the words with low confidence ratings so that users can selectively rely on the opaque keywords. Our work is the first to incorporate the semantic structuring of speech in the document annotation context. The new challenge was to display the caption interface near the anchor text without occluding underlying text. As a solution, we embraced the speech interface in the flow of body text by leveraging the fluid document layout technique.

Lastly, when audio is recorded in conjunction with other timestamped data, such as digital ink, it is also possible to navigate the audio using the linked data. For example, ink strokes recorded along with voice can be used to skip to the voice stream corresponding to a particular stroke (Stifelman et al., 2001; Whittaker et al., 1994; Wilcox et al., 1997). Our work introduced a variation of this approach by introducing gesture traces (Yoon et al., 2014) as an additional type of navigation cue.

This section reviewed a series of previous studies that identified and addressed the problems of speech on the recipient side. What follows are the problems on the other side of communication, who records and edits speech.

2.2.2 PRODUCING SPEECH[2]

The production issue of speech commenting has been frequently documented in previous studies (Hew & Cheung, 2013; Marriott & Hiscock, 2002; Scholl et al.,

[2] The text of this section was derived from a previous publication (Arawjo et al., 2017).

2006; Sivaraman et al., 2016). Overall, there were two types of problems: high workload of editing recorded speech, and mental burdens of recording live voice.

Editing recorded speech is a tedious and time-consuming job in comparison with text-editing. Through the years, different designs to speech editing interfaces have been proposed and evolved. In professional audio/video production software, low-level audio editing was made possible through waveform representation over a timeline (Adobe, 2016b, 2016c; Apple, 2016; Audacity Team, 2016). Since the fine-grained timeline in the waveform interface introduced an extra burden on novice users, especially in the context of speech editing, researchers structured audio into a higher-level representation by analyzing its acoustic structure, such as speech/non-speech chunks, phrases, or sentences, for browsing (Arons, 1993; Stifelman et al., 2001) or editing (Ades & Swinehart, 1986; Hindus & Schmandt, 1992). A more recent approach employs time-synchronized captions from automatic speech recognition to augment the edited audio chunk with semantic meaning. With this approach, snapping the editing prompt to a word's boundary afforded text-like audio/video editing (Berthouzoz et al., 2012; Rubin et al., 2013; Sivaraman et al., 2016; Whittaker & Amento, 2004). Contrary to the previous transcription-based systems where audio is loosely coupled with the transcription, synthesized audio of our TypeTalker system is guaranteed to say exactly what is written and edited as text. This approach not only enables faster speech revision, but also supports generative editing operations, such as insertion or rephrasing of words that used to require cumbersome rerecording through keyboard input.

Previous research on speech interfaces found design factors that affect speaker psychology. While ordinary voice conversation is ephemeral, if speech interfaces give people a sense of being recorded, they tend to speak differently (Clark, 1996). Even a physical awareness about the recording device (e.g., awareness about a wired lapel microphone) can impact speaker response, reducing creativity and disclosure and introducing speech disfluencies (Wang & Nass, 2005). Recent studies on asynchronous speech commenting systems suggest several factors that might increase cognitive load for speech commenting: concerns about one's speech disfluencies, affective disturbance of hearing one's own voice, and lack of lightweight editing features (Holzman & Rousey, 1966; Marriott & Hiscock, 2002). In this work, we presented a solution that replaces user voice with the synthesized machine's voice. The results of our evaluation show that this synthesis-based approach reduced speech anxiety and self-affective disturbance of the speaker.

2.3 Deictic gestures

Communicative bodily movements, gestures, are the crux of nonverbal communication in face-to-face encounters. Use of collocated gestures was widely and extensively observed in different collaboration settings including collaborative design work (Bekker et al., 1995; Tang, 1991), instructed machine operation (Kuzuoka, 1992), and document explanation (Bekker et al., 1995; Bickmore et al., 2008; Cox & Cox, 2008; Whittaker et al., 1993).

There are many different gesture taxonomies based on varying perspectives, but the category of deictic gesture has always stood out as unique and significant (Bekker

et al., 1995; Clark, 1996; Karam & Schraefel, 2005; McNeill, 1992; McNeill, 2005). The commonality of the examples of deictic gesture in the literature—pointing at, tapping on, or waving over a target—is that they carry indexical information that brings an external referent, whether it is physical or abstract, into the conversational context. By focusing on the ambivalent and equivocal nature of deictic gesture, McNeill (2005) claimed that a gesture does not fall into a single exclusive category, but contains different extents of saliency in each of the different functional dimensions, where *deixis* takes one axis. In other words, any gesture can be regarded as a deictic gesture if it has high deictic saliency.

A pointing gesture, however, is still worth highlighting as a primary and special type of deictic gesture. There are abundant studies on the primacy of pointing. Pointing can be universally understood without cultural convention (McNeill, 2005). It is so fundamental and elementary that even infants and chimpanzees can learn, understand, and use pointing (Blurton-Jones, 1972; Kita, 2003; Ong, 1989). Direct visual intuition is another reason why pointing is so popular. Projecting a hypothetical 3D ray from one's index finger to the target—commonly referred to as G-shape gesture—raises vivid imagery of spatial indexing (Kita, 2003). In this regard, it is not surprising that the pointing gesture is the most popular deictic gesture in the literature, and moreover, some studies even use the term "pointing" almost synonymously with deictic gesture (Clark, 1996; Karam & Schraefel, 2005; Kita, 2003).

Previous studies in HCI have examined the deictic role of gestures in online communications. Boom Chameleon allowed users to highlight a point of a 3D scene using glowing blobs that are recorded in sync with voice (Tsang et al., 2002). Fussell

et al. (2004) transmitted and overlaid pen-drawing traces over remote live-video feeds for a deictic purposes. Harrison et al.'s (1999) electronic cocktail napkin system could capture a remote user's hand gesture over an upward facing monitor as a video stream recorded by a downward facing camera. Lee and Tatar (2012) comparatively evaluated different deictic markups as a visual aid for collaborative Sudoku puzzles. In contrast, our work focuses on the use of gesture in *asynchronous document annotation* where recorded pointing gestures are animated over written documents. As such, our study provides added insights about the way voice and gesture annotations are created and replayed in the context of the underlying text.[3]

2.4 Multimodal annotation as an educational technology[4]

There is a long tradition of leveraging multimodality in various types of educational settings. The first attempt to take advantage of recorded voice in the classroom was when instructors recorded their feedback using a cassette tape recorder (Anson, 1997; Klammer, 1973). Later, instructors recorded video of themselves editing students' papers (Crook et al., 2012; Silva, 2012) or provided anchored audio comments over a PDF document (Oomen-Early et al., 2008). Online discussion forums now support threaded voice comments to facilitate discussions between students (Hew & Cheung, 2012; Marriott & Hiscock, 2002). Our work goes beyond

[3] The text of this section was derived from a previous publication (Yoon et al., 2016).

[4] The text of this section was derived from a previous publication (Yoon et al., 2016).

these previous efforts by (1) designing a new educational system that leverages communicative capacities of new hardware and interaction techniques, and (2) investigating student perceptions on the use of rich commenting for instructor feedback and peer discussion purposes.

Voice communication has been demonstrated to improve student–student and student–instructor engagement as well as give a sense of the instructor's social presence (Ice et al., 2007; Oomen-Early et al., 2008; Tu & McIsaac, 2002). Mayer's (2005) work on multimedia learning indicates that audio communication is particularly useful in situations where an emotional connection between speaker and listener is desirable. This dissertation, by deploying our thesis system to real classrooms, offers concrete evidence about how and why voice can accommodate the relationship between instructor and students.

3 TextTearing: Expanding whitespace for digital ink annotation[5]

Written annotations are a crucial aspect of engaged reading activities. Among the many roles that annotations play include demarking important passages, tracking reading progress, and recording interpretations about the text (Marshall, 1997). Paper documents provide a number of affordances that cater to annotation activities (Sellen & Harper, 2003). For instance, blank spaces in the document—margins, within-text spaces, and blank pages—offer readily available regions where annotations can be created with minimal interruption to the reading process. Moreover, the spatial proximity of these regions to the text provides context (Golovchinsky et al., 1999) and implicitly connects annotations with the text to which they refer (Marshall, 1998). Paper materials are not perfect, however. Pages have a set amount of space for markup, which can be insufficient at times (Pearson et al., 2009).

Support for annotation is less robust when it comes to digital documents. Although many software tools support free-form inking with touchscreens or pen digitizers, writing on electronic screens tends to place more demand on the available annotation space (Agrawala & Shilman, 2005). However, the dynamic nature of digital documents provides an avenue for workarounds. Existing strategies such as comment boxes, digital Post-it™ notes (Pearson et al., 2011), and the ability to insert blank

[5] The text and figures of this chapter were derived from a previous publication (Yoon et al., 2013).

pages expand the space for annotations. Unfortunately, these come at the cost of fluidity and can disrupt the reading process (Sellen & Harper, 2003). Moreover, displaying these types of annotations (in overlaid windows alongside the text, for example) can result in the main text being obscured or unpredictable annotation layouts.

We introduce an interaction called TextTearing that addresses these problems. With this technique, users can tear open (i.e., expand) the whitespace between adjacent lines of text. This allows users to create blank space where it is needed, while maintaining the overall logical structure of the text and avoiding occlusions. We describe an initial design of our system, which includes our approach for rearranging document elements, a two-handed tearing interaction based on Zeleznik et al.'s work (Zeleznik et al., 2010), and a palm rejection system that makes it possible to implement this technique on commodity hardware. A complementary technique for expanding the page margins is also described.

Our evaluation of TextTearing compares it against a baseline where there is ample whitespace to directly write, a condition with expandable side margins, a two-handed version using an alternative tearing gesture, and a pen-only tearing technique based on pigtail gestures (Hinckley et al., 2005). Our results showed that margin expansion was comparable to the baseline conditions, which were fastest because they required no interaction. However, the two highest ranked techniques in terms of user preferences employed tearing. In contrast, the baseline condition came in last, suggesting that the placement of the writing space makes a difference. Among the tearing techniques, the

pen-only pigtail technique had a preference advantage over our initial design and could potentially be faster in practice.

3.1 TextTearing

Spatial proximity between annotations and the primary text provides context (Marshall, 1997), helps maintain continuity of attention, and serves as an implicit connection between the annotation and the text (Marshall, 1998). Thus, the central goal of our system is to give readers access to writing space anywhere beside printed text. To accomplish this, we let users create an expandable region of whitespace by adjusting the spacing between lines of text (see Figure 1). As the region grows, the content below it is shifted lower in the page. In the following sections, we first describe the algorithms we use to make the system applicable to a wider variety documents; specifically, those with multiple columns. Then, we present the interaction for performing the space expansion.

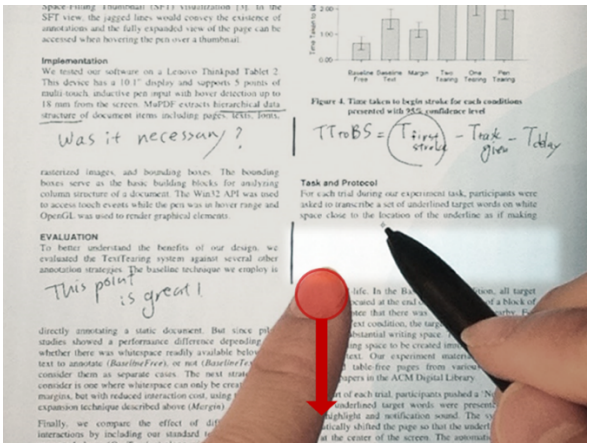


Figure 1. The TextTearing system allows additional writing space to be created between lines of text using a downward tearing gesture. The expanding region is highlighted. The red arrow indicates touch gesture opening a space. The blue shadow is a dynamic palm rejection region.

3.1.1 HANDLING MULTI-COLUMN LAYOUT

To reduce undesirable effects of layout modification, such as introducing spurious spaces in adjacent columns, or collisions of shifted text into other elements, it is important for the system to understand the visual structure of the document and to enforce some modification constraints.

Our first task is to identify the position of text columns. Since our system targets PDF documents, we leverage layout information about the start and end points of the lines of text on a page. We first remove short lines of text from our analysis. Then, we use the midpoint between minimum and maximum horizontal line positions to identify the position of the alley between columns. Once that is done, we assign each line and figure into either of the left column, the right column, or leave it alone if it doesn't fall cleanly on either side. Once lines have been classified into the columns, the unclassified text is placed into either the header or footer depending on whether they are above or below the column bounding boxes. A typical structure our algorithm produces is shown in Figure 2. We currently only support two-column page layouts but this algorithm can be easily extended to handle layouts with more than two well-defined columns.

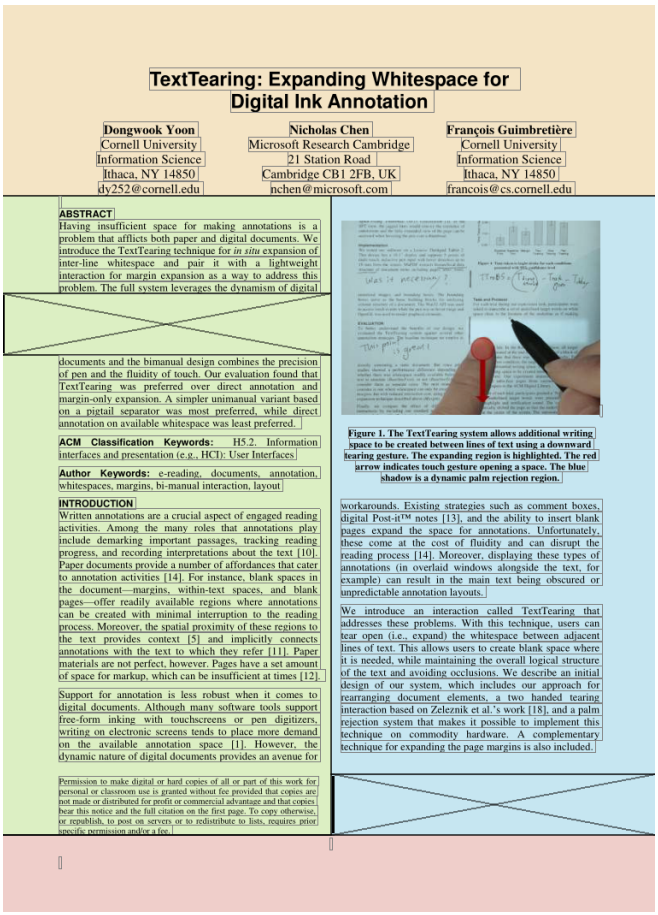


Figure 2. The dynamic document layout model. Boxes with an "X" are movable regions of the document. The area above and below the columns are each a single block. The left column has two text block and one region produced using TextTearing; the right, one text blocks and one tearing region.

The extracted document structure dictates how the document expands. Tearing operations will always maintain the relative relationships between the header, columns, and footer. For instance, if space is created in one column, then the text below it will be pushed downward, which will in turn push the footer downward, creating whitespace at the bottom of the adjacent column if needed (Figure 2). Enforcing this high-level structure isolates expansion to a single column, which results in more predictable behavior and helps preserve spatial relationships in the document.

In this first prototype we did not consider the case in which a tearing operation crosses over a long stroke such as the line in a callout, for example. This could be addressed by augmenting our system with the annotation reflow technique proposed by Golovchinsky and Denoue (Golovchinsky & Denoue, 2002).

Annotations are anchored to neighboring pieces of text to accommodate subsequent movements of that text. This is done by first grouping strokes that are created in rapid succession (< 0.5s separation interval) together and then setting the stroke coordinates to be relative to the location of the closest piece of text so that when the text moves, so do the strokes.

3.1.2 TEARING INTERACTION

A guiding principle for the tearing interaction is that it should be precise (since lines could be closely spaced) and introduce minimal overhead compared to ordinary inking. Hinckley et al. (2010, 2012) showed that pen + touch input combines the lightweight nature of touch with the accuracy of pen input. For this reason, we adapt the pen + touch approach Zeleznik et al. (2010) employed to create blank space in Hands-on-Math. In our system, users create additional space by hovering the pen tip over a target location and then use the finger on their other hand to "tear" the surrounding text apart. To avoid false positives that can occur when users are panning with the pen near the screen, we require that the initial finger touch must be within 16.5 mm of the line of text the pen is hovering over. Since we do not overload the hovering state, it is not necessary to have a feed forward widget like the one in Hands-on-Math.

Dynamic Palm Rejection

To provide pen + touch input along with a comfortable writing environment, we needed some way to reject touch events generated when the palm of the hand holding the pen inadvertently comes in contact with the screen. We noticed that Vogel et al.'s occlusion silhouettes model (Vogel & Balakrishnan, 2010) provided a conservative estimate of the projection of the hand on the screen. Therefore, this model can be used to compute the area where the palm could possibly make contact with the screen. The palm rejection system can then reject any contact points in this area. Remaining touch points are passed to our application to be used for the tearing interaction as well as for panning and zooming when the pen is over the screen.

Margin Expansion

The tearing interaction described above could conceivably be used for expanding page margins as well. We discovered, however, that it was more straightforward to simply expand the margins when users pan past the current page extents. The expansion behavior does not interfere with page turning since our system uses a side tap to turn pages. Support for page turning with swipe gestures can be achieved using either a swipe starting on the bezel or a threshold mechanism to differentiate panning from swiping. Although we only enable this capability for the horizontal margins, this simplified interaction could be used for vertical margins as well.

Viewing the Original Document Layout

A three-finger touch gesture shows an alternative view of the document in which the dynamically generated whitespace regions are collapsed. This collapsed view of

the document can be useful for checking the original document appearance or for ensuring that a full page can be displayed on the screen.

In the collapsed view, each collapsed area is represented by a jagged underline that reflects the projection of ink strokes onto the tear line. While maintaining this quasi-mode, hovering the pen over a jagged line overlays the associated annotation region over the page in a semi-transparent window. We also prototyped a "spring loaded" version of the software that collapses newly created regions by default. In this alternative design, keeping the pen tip in hover range keeps the newly expanded region open, and exiting the range restores the original document layout. Our experience with the technique suggested that a more effortless and stable way to maintain the quasi-mode is needed, however.

The collapsed view of the document may also be useful for maintaining a constant page aspect ratio for alternative visualizations such as Space-Filling Thumbnails (SFT) (Cockburn et al., 2006). In the SFT view, the jagged lines would continue to convey the existence of annotations and the fully expanded view of the page can be accessed when hovering the pen over a thumbnail.

Implementation

We tested our software on a Lenovo Thinkpad Tablet 2. This device has a 10.1" display, multi-touch input supporting up to 5 simultaneous points of contact, and inductive pen input with hover detection up to 18 mm from the screen. We used the MuPDF library in conjunction with OpenGL to render PDF documents onto the screen. MuPDF provided the hierarchical data structure of document items including

page boundaries, text, raster images, and bounding boxes. The bounding boxes served as the basic building blocks for analyzing the column structure of a document. The palm rejection scheme currently used in Windows 8 does not support simultaneous bimanual interaction as it disables all touch input events when the pen is in hover range. We were able to work around this issue by using the RawInputDevice Win32 API to access touch events while the pen was in hover range.

3.2 Evaluation

To better understand the benefits of our design, we evaluated the TextTearing system against several other annotation strategies. The baseline technique we employ is directly annotating a static document. But since pilot studies showed a performance difference depending on whether there was whitespace readily available below the text to annotate (*BaselineFree*), or not (*BaselineText*), we consider them as separate cases. The next strategy we consider is one where whitespace can only be created in the margins, but with reduced interaction cost, using the margin expansion technique described above (Margin).

Finally, we compare the effect of different tearing interactions by including our standard tearing technique presented above (*OneTearing*) along with two alternatives (Figure 3). The first alternative (*TwoTearing*), replaces the single finger non-dominant hand (NDH) interaction with a two finger NDH interaction resembling the gesture used for zooming on multi-touch devices. While hovering the pen, spreading one's fingers from a pinching position increases the size of the whitespace region. A possible benefit of this approach is that the positions of the NDH fingers map directly

to the expansion region. The second alternative (*PenTearing*) uses a single handed pen interaction based on pigtail delimiters (Hinckley et al., 2005). In this condition, a long horizontal line (> 12.0 mm) followed by a pigtail starts the tearing operation. The centroid of the stroke points preceding the crossing point indicates where the whitespace expansion should occur. The pen stroke is then extended downwards to specify the size of the expansion. The trigger is distinctive enough that it does not seem to interfere with writing activity. We considered but ruled out using the pen barrel button as a trigger since it can be error-prone (Song et al., 2011).

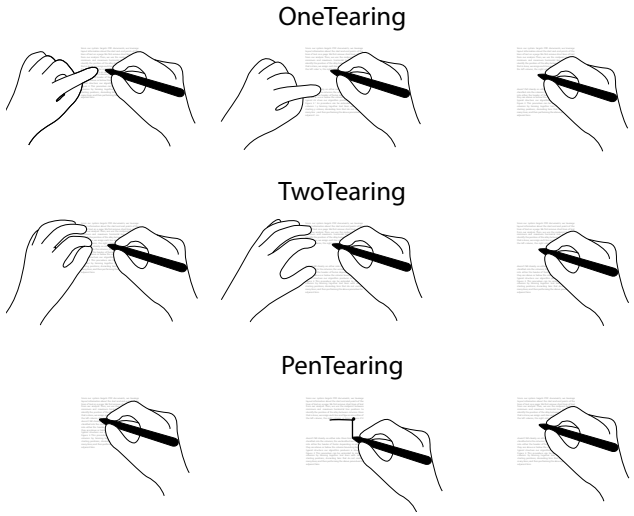


Figure 3. Tearing gestures considered in the experiment.

Task and Protocol

For each trial during our experiment task, participants were asked to transcribe a set of underlined target words on white space close to the location of the underline as if making notes in real-life. In the BaselineFree condition, all target words were located at the end or the beginning of a block of text to guarantee that there was writing space nearby. For the BaselineText condition, the target was at least four lines

away from substantial writing space. The tearing conditions required writing space to be created immediately below the underlined text. Our experiment materials consisted of picture and table-free pages from various two-column formatted papers in the ACM Digital Library.

At the start of each trial, participants pushed a 'Next' button and the underlined target words were presented with a visual highlight and notification sound. The system then automatically shifted the page so that the underlined words were at the center of the screen. The automatic centering removed page adjustment performance as a factor so that the measured time better reflects the amount of time it takes the participant to find or create writing space. Our implementation required a one-time iterative calibration procedure to manually tune the 5 parameters of the hand occlusion model (e.g., hand radius, forearm angle, etc.). This process could be made automatic, however (Vogel & Balakrishnan, 2010).

Each technique was tested in one block. A block consisted of an introduction to the technique, a practice phase with 18 trials, and an experimental phase with 13 trials. The order of the techniques and materials was counterbalanced using a 6×6 Latin square. At the end of each block, participants were interviewed and completed a paper-based NASA Task Load Index survey (TLX). After all blocks were finished, the participant sorted 5 cards, representing the baseline and the four other techniques we tested, by preference. There were 12 participants in total (10 females, 2 males, average age 23.2 years old). Participants were paid $10 for the hour-long experiment.

3.3 Results

Measurements of the elapsed time were computed starting from when the underlined target words are shown to the beginning of the first inking stroke. We accounted for the skewed distribution of human reaction time by taking the median value of a block's trials. We used Greenhouse-Geisser correction when we could not assume sphericity and the Bonferroni correction for pairwise comparisons.

Errors

We counted an error whenever participants opened a space below a line other than the one that was underlined. A one-way repeated measure ANOVA on error rate showed a marginally significant effect of technique ($F(1.56, 17.17) = 3.73$, $p < .054$, partial $\eta 2 = .25$). TwoTearing ($M = 13.69\%$, $SD = .044$) showed a significantly higher error rate than PenTearing ($M = 2.78\%$, $SD = .054$, $p < .05$), due to small panning motions induced by the first finger prior to the second finger touching down. OneTearing fell in the middle ($M = 7.63\%$, $SD = .024$) but the differences were not significant. We also measured the occurrence of false tearing activations for the pigtail gesture and found a single case across all trials (0.6%) that occurred when the participant was writing in cursive.

Total Space Opening Time

A one-way repeated measure ANOVA on the total space opening time showed a significant effect of technique ($F(2.47, 27.18) = 19.5$, $p < .001$, partial $\eta^2 = .64$). BaselineFree was the fastest condition ($M = .65$ sec, $SD = .4$, $p < .004$) followed by the Margin condition ($M = 1.19$ sec, $SD = .48$) that was marginally faster than

BaselineText (M = 1.60 sec, SD = .60, $p < .067$) with both significantly faster than all tearing techniques ($p < .037$). It was somewhat surprising that the Margin technique performed in between the two baseline conditions that required no interaction at all. We believe that even though extra time was required to perform the margin expansion it is offset by the savings from not having to find a place to write.

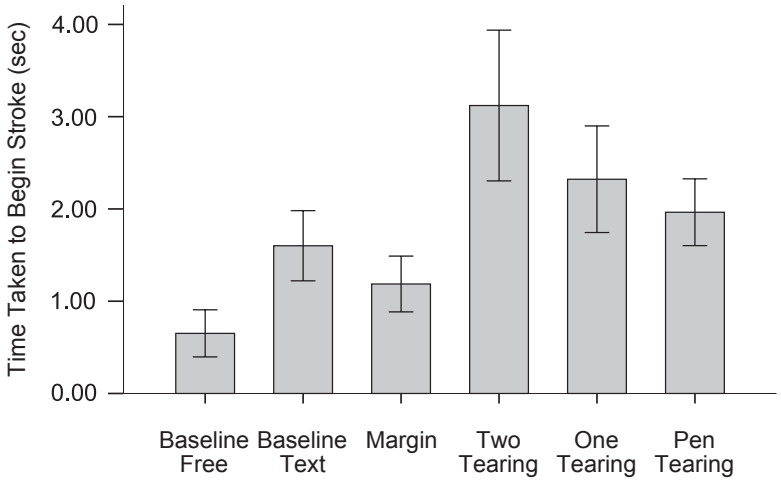


Figure 4. Time taken to begin stroke for each condition presented with 95% confidence level.

For the tearing-enabled interactions, TwoTearing was the slowest (M = 3.12 sec, SD = .37) but the difference was only significant versus PenTearing (M = 1.96 sec, SD = .57, $p < .005$), but not OneTearing (M = 2.32 sec, SD = .91, $p < .098$). Given the large difference in means, we looked more carefully at our data and discovered that an overly cautious participant caused a large increase in the variance. Re-running the analysis without this participant resulted in TwoTearing being significantly slower than both of the other techniques ($p < .005$). The difference between PenTrearing and OneTearing was not significant ($p < .18$). To better characterize differences between OneTearing and PenTearing, we decomposed the total space opening time into 3 components: the time it took to begin tearing (T^{begin}), the time for the tearing action

(T^{Tear}) and the time it took to start writing after the tearing action was completed ($T^{1st\ stroke}$). We found that most of the difference between the two techniques were from T^{Tear}, with PenTearing being faster ($t(11) = 5.43$, $p < .0001$).

Subjective Preferences

We performed a Friedman test on the ranking data we gathered at the end of each session. We found a significant difference in ranking ($\chi 2 = 46.80$, $p < .001$), with PenTearing (1.25) coming first followed by OneTearing (2.25), Margin (3.17), and TwoTearing (3.75). The baseline technique was the last (4.58). The TLX showed that TwoTearing required the highest effort ($F(2.81, 30.87) = 5.78$, $p < .003$, partial $\eta^2 = .34$). No other significant differences were observed.

3.4 Discussion

PenTearing was the most preferred, and rated higher than the Baseline techniques even though it was 3 times slower than direct annotations. We believe this effect illustrates that directly writing in naturally occurring space is not necessarily desirable even though it can be done quickly. This interpretation is in line with Agrawala and Shilman's observations about the difficulties of writing on electronic documents (Agrawala & Shilman, 2005). Additional studies are needed to determine how our expansion-based techniques compare to a zooming scheme like DIZI.

The fact that two of the tearing techniques were preferred over margin annotation reinforces the notion that the spatial positioning of the annotation is important. One participant noted that creating a space below the target text for writing resulted in less occlusion from the writing hand compared to annotating beside the target text (i.e.

writing in the margin). We believe that reduced occlusion, along with ergonomic and semantic benefits from being able to position annotations flexibly, are advantages that TextTearing would bring to real-world annotation tasks.

Participants' preference for the one-handed, pen-only, tearing technique over the two-handed ones was surprising given the expected advantages of bimanual interaction (Guiard, 1987). Participants told us that PenTearing was a simpler one-step gesture. They also pointed out that without hover as a mode delimiter, PenTearing allowed the pen tip to remain on the screen, which resulted in more stable targeting. Finally, participants mentioned that performing the tearing interaction using a single hand produced less occlusion than using two hands and freed the ND hand for other, unrelated, tasks. From an implementation standpoint, the simpler input hardware required is an additional advantage of the one-handed technique. We believe, however, that additional investigations of bimanual versus unimanual modalities in the context of real-world activities are needed to definitively confirm these advantages.

Palm Rejection

Despite the fact that users were required to have the pen in the hover zone before resting their palm on the screen, no participant thought avoiding these false-positives was burdensome. One situation where the occlusion model failed was when users rotated the screen prior to writing, such as to squeeze text into a narrow space during a baseline interaction. This situation could potentially be remedied by using on-board motion sensors. It bears noting that the palm rejection functionality is still useful even

with a technique like PenTearing because it enables other touch gestures, such as navigation, while the pen is in the hover volume.

3.5 Summary

TextTearing interaction addresses the problem of lack of space for written annotations. We leveraged the fluid nature of a digital document to dynamically restructure column layout to interleave new whitespace as a part of the page. We designed, implemented, and evaluated several versions of TextTearing gestures that combined pen and touch inputs. Among them, the pen-only pigtail gesture was the most preferred because of its speed and flexibility.

4 RichReview: A multimodal annotation system[6]

This chapter focuses on our endeavor to expand expressivity of annotations by introducing rich interaction modes on top of stylus writing. Although a variety of HCI studies have presented types of multimodal annotation systems (Levine & Ehrlich, 1991; Marriott & Hiscock, 2002; Neuwirth et al., 1994; Tsang et al., 2002), a persistent problem is that none of them successfully balances expressiveness of the system and accompanying complexity of user interfaces. For example, speech commenting requires complex audio navigation panels that clutter the interface. This leads on to the goal of this dissertation, that is, to build an annotation system that offers unmatched expressivity to compete with the face-to-face meeting while making its interface as simple as that of textual annotation tools.

With this goal in mind, we built RichReview, a multimodal annotation system that simultaneously records multiple aspects of users' communicative expressions and replays them in a synchronized fashion at remote locations. For instance, as shown in Figure 5, a commenter can record speech to provide verbal descriptions while creating digital ink markups and hovering the stylus over the screen to point and refer to different parts of the page. On the recipient side, these multimedia components are replayed in synchronicity, thereby delivering vivid and engaging sensations as if the

[6] The text and figures of this chapter were derived from a previous publication (Yoon et al., 2014).

commenter is narrating over, writing on, and pointing at the page. The rest of this chapter covers the details of our design principles, strategies, and evaluations.

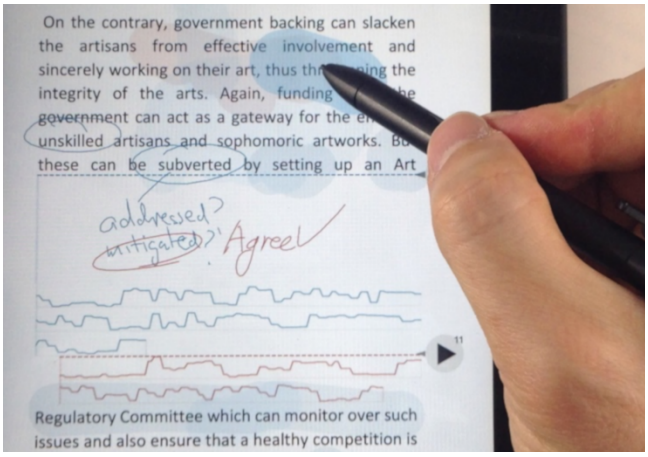


Figure 5. RichReview running on a tablet. Hovering the pen over the screen leaves traces of gesture (blue blob on top). Inking can be done on expanded space (middle). Voice recording is shown as waveform (bottom).

4.1 Design principles

The literatures in the chapter 2.1 and 2.2 are indicative of the design challenges for developing an effective multimodal collaborative annotation system. With that in mind, we proposed the following principles:

Limiting system complexity

Introducing multiple annotation modalities runs the risk of bringing in additional complexity and overhead. The added overhead could then affect all annotation activities. Therefore, a significant part of the design of our system was focused on ensuring that annotations are created and consumed in ways that are lightweight and fluid. Satisfying this design goal argued against locking the user into interaction

modes or adding additional interaction steps. RichReview employs a simple and consistent set of interactions for creating any kind of annotation.

Versatility and Choice

The literature provides many examples showing that the optimal modality for communicating varies by its content and purpose. For example, inking is popular for lightweight copyediting, and a combination of voice and pointing can be useful to describe structural issues of a writing. For this reason, a second design goal is to allow users to employ a flexible mix of annotation modalities.

Balancing Emphasis on Production and Consumption

The success of groupware is contingent on the balance of benefits to different stakeholders (Grudin, 1988). Thus, an annotation system that supports collaborative tasks must focus as much on improving the ability of recipients to skim, access, and revisit annotations as it does on supporting the creation of them in the first place. Given that non-textual content can be difficult to access and skim, we place an emphasis on techniques that assist users in consuming rich annotations.

The first version of RichReview was designed for use with tablet devices, since tablets are a preferred form factor for active reading activities (Morris; et al., 2007). Moreover, current tablet devices contain the necessary hardware to capture the three input modalities we are interested in.

4.2 Creating multimodal annotation

In designing the comment creating features of RichReview, we paid special attention to keep the interface as simple and direct as possible. This means that our design minimized the number of interaction modes and the number of interaction steps to get the tasks done. For example, RichReview retains the paper metaphor in which inking can be performed anytime without entering a special mode. Also, the inking can be done in any visible space (e.g., including whitespace, text, even on the other annotation system. See Figure 6.), as we designed the inking operation to be oblivious to the type of underlying surfaces,

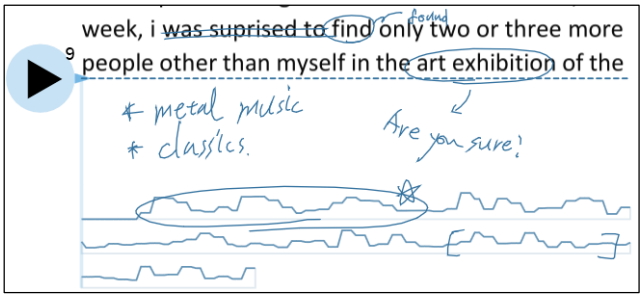


Figure 6. A mixture of static ink annotations along with the playback control for a multimodal annotation.

RichReview's multimodal annotation recordings capture voice in conjunction with ink and pointing gestures. RichReview requires users to explicitly start and stop the recordings to dispel privacy concerns associated with a system that is always-on. A recording session is started using an underline followed by a pigtail that extends in the horizontal direction (Figure 7 (a)). This gesture was selected due to its similarity to the best performing gesture for TextTearing in order to reinforce the idea that annotation activities commence with an underline followed by a pigtail. The location of the

underline specifies the anchor point of the annotation and creates a small playback control icon in the margins at the same vertical position on the page. The icon doubles as a marking menu containing commands for working with the annotation.

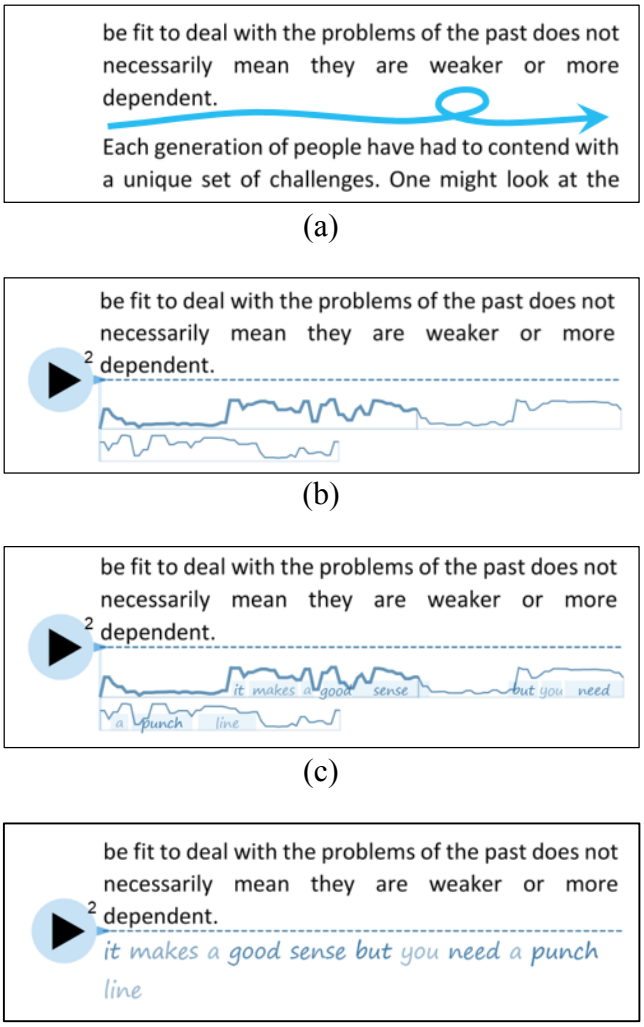


Figure 7. Recording and visualizing speech: (a) Pigtail gesture to begin and anchor recording (b) Waveform during replay (c) Waveform with word overlays (d) Transcription with varying opacities based on recognition confidence.

Capturing Voice

When the recording session begins, a small amount of extra space is inserted between the lines of text where the initial underline gesture is drawn. Inside this space, a waveform representation of the captured audio grows from left to right. Upon reaching the end of the line, the space expands slightly downward and the waveform continues into the new space.

RichReview also includes features to help users add structure to their speech annotations. Similar to Audio Notebook (Stifelman et al., 2001), users can structure their voice annotations by creating time-indexed ink notes that the recipient can later use to jump into import parts of the annotation. Also, performing an annotation creation gesture while a recording session is active ends the active recording and immediately starts a new one. This is useful when the annotation moves to a different topic; the interaction saves the user from the interruption incurred from stopping the recording and creating a new one. Performing the annotation creation gesture over an existing annotation appends a new recording to the existing one.

Capturing Pointing Gesture

Pointing at a location in a document is a fast and lightweight way of supporting discussion with reference to specific parts of a document (Bickmore et al., 2008). RichReview provides the Spotlight interaction to reproduce this capability. With the Spotlight interaction, hovering the pen over the page while recording creates a circular translucent region at the pen's position (Figure 5, page 59). When recording ends, translucent trails of where the Spotlight has been are shown on the document.

Another way that a user can communicate the location of the region of interest is through the creator's viewpoint (i.e. what the creator was looking at during the recording). Similar cues are used in F2F collaboration by observing a collaborator's gaze. To convey this information to recipients, RichReview records viewpoint adjustment operations such as panning and pinch-to-zoom gestures for later playback.

Audio Post-Processing

When a recording is complete, the captured audio is passed to an automatic speech recognizer running in the background. When the transcription is complete, the words in the transcript are shown over the portion of the waveforms corresponding to when they were spoken (Figure 7 (c)). Displaying words over the existing waveform maintains visual continuity with unadorned waveform representation. However, when improved readability of the transcript is desirable, users can display the transcription on its own (Figure 7 (d)) by selecting an option from the marking menu.

The transcribed audio can be used to trim or tidy up the audio in the audio editing tool (Figure 8). In the editing tool, crossing through words or portions of the waveform or transcript grays out those sections of the recording and removes them from the recording. Crossing through a deleted section reverses the deletion. Edits made with the tool are automatically snapped to word boundaries so that the result does not slice the audio in the middle of a word.

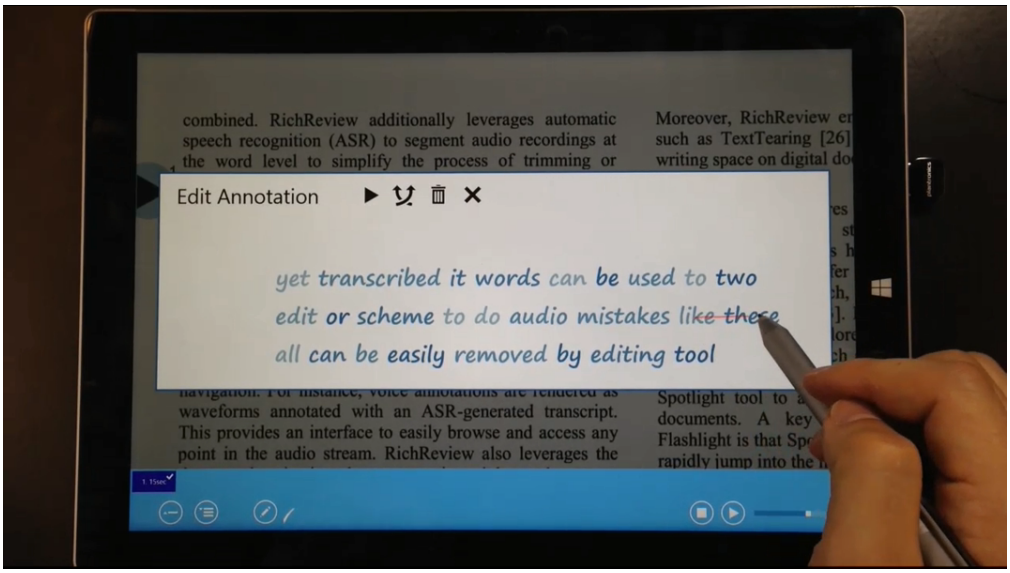


Figure 8. The transcription-based audio editing interface of the RichReview tablet app.

4.3 Consuming multimodal annotations

The ink, gesture, and audio (and transcript) within a recording share the same timebase. RichReview leverages time synchronization to provide a rich rendering of the way the original annotation was created and lets users quickly jump to a specific part of the annotation stream.

4.3.1 BASIC PLAYBACK

Basic access to annotation recordings is through the play icon at the annotation anchor or media control at the bottom of the screen (Figure 9). During playback, ink is rendered in a grayed out form if playback has not reached the point where it was created (Figure 10) and then drawn with a colored stroke afterwards. Spotlight traces are rendered as an animated, translucent circle.

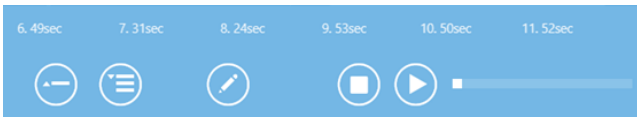


Figure 9. Recording list and media control. A list of annotations, sorted in order of creation, runs across the top. Buttons are used to collapse, expand, edit, stop and play annotations, respectively.

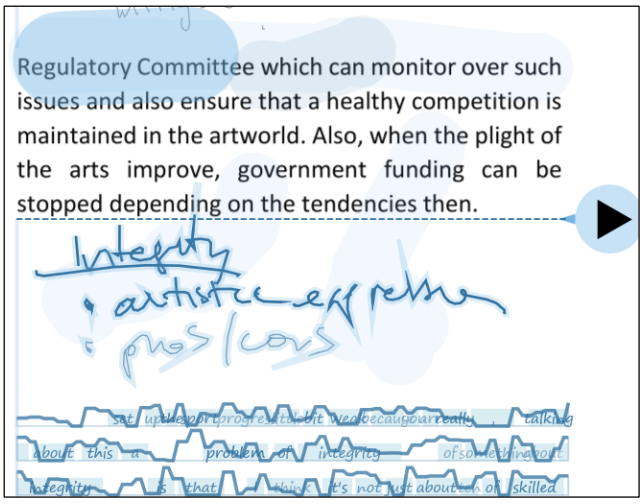


Figure 10. Spotlight trails along with dynamic ink. Recorded strokes are dynamically replayed as playback advances. Grayed out strokes will come in the future. The speech annotation here is structured by writing keywords while speaking.

4.3.2 CROSS-MODAL INDEXING FOR ENHANCED NAVIGATION OF MULTIMODAL ANNOTATIONS

Although the basic playback controls are sufficient for the linear consumption of annotation content, they can be inadequate for random access. For example, users may wish to skim through annotations or visit a specific part of an annotation. RichReview offers several features that support these more complex navigation tasks.

For example, users can tap on a point in the waveform or transcript to skip to the corresponding point in the annotation recording. The waveform can be useful for finding gaps in the audio, which often delimit sections within an annotation stream.

For finer-grained navigation, the transcript (Figure 7 (d)) can be used to visit parts of an annotation based on words of interest. The need for random-access to audio is critical in light of the fact that speech-to-text technology can still be quite error-prone; generally, it is not possible to use the transcript on its own to consume speech content. Therefore, it is imperative that users have a way to quickly jump to and listen to the actual audio.

Ink strokes and Spotlight trails can similarly be used to index into an annotation. One important design decision we made was to show the entirety of the ink and Spotlight traces at all times, so that they can be promptly accessed when needed. On the one hand, this choice does not preserve the exact appearance of the page during annotation creation. On the other hand, we believed that giving the ability to skip forwards into an annotation outweighed this concern. We distinguish between strokes that have been made and those that have yet to appear by rendering strokes in different colors.

We also explored other ways of leveraging the links between different modalities that were not as useful. For example, in early prototypes, we highlighted portions of the waveform if they corresponded to times when inking or Spotlight was active. We found that these highlights were not very useful because the highlights provided few cues about the specific objects on the page to which they referred.

4.3.3 CREATING CONVERSATIONAL THREADS

Given the iterative nature of collaborative writing tasks, RichReview provides collaborative annotation features that allow users to respond to existing annotations

made by peers. These features help “close the loop” when people collaborate on a document. In RichReview, annotation entities, such as waveforms, ink, playback controls, and Spotlight traces are color-coded by user identity. However, tracking and showing user identity is only a small part of providing multi-user support.

RichReview differs from other collaborative annotation systems in that it is possible to respond to an annotation using a different modality. The way RichReview enables this is to treat ink and audio annotations in the same way as the underlying body text of the document. There are two benefits of this design decision.

Figure 11. Red user inserted a voice annotation in the middle of existing Blue user’s voice transcript (footage from the real-user data, P7). Red user’s Spotlight is anchored on the transcript.

First, mark-up operations that could be applied to the original document can also be applied to annotation entities. For example, Spotlight can be used to bring attention

to a part of the transcript (Figure 11) and ink can be used to circle portions of a waveform (Figure 12).

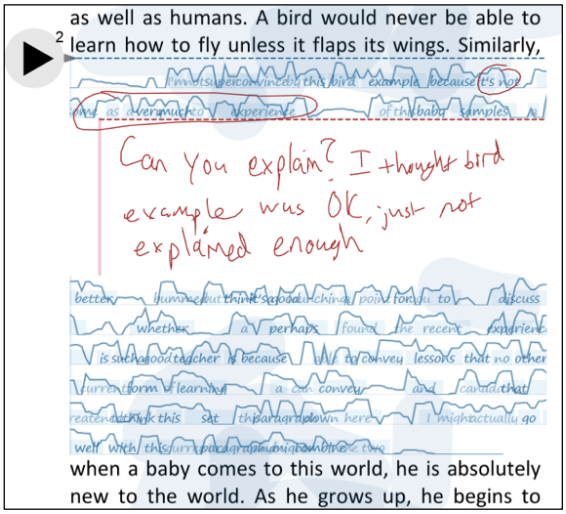


Figure 12. Circling on waveform to designate a part of the voice (footage from the real-user data, P10)

Second, treating annotations like the underlying text provides multimodal support for discussions between collaborators conducted through threaded comments. For instance, users can create an expansion space in the middle of an audio waveform and insert a comment using ink (Figure 12). Inserting a voice comment under an existing inking space is also possible (Figure 13, top).

In some cases, interleaving annotations with the body text can break the flow of reading due to annotations having a large visual footprint. In these situations, RichReview allows users to collapse or expand annotations. An option in the marking menu (Kurtenbach, 1993) accessed through the play icon allows this to be done on a per-annotation basis. Collapse-all and expand-all buttons in the bottom toolbar can also be used to switch back and forth between the original document layout and the

fluid layout. These multi-user features that allow users to converse and engage in discussion through annotations are illustrative of how our initial design goals of interactional consistency and flexibility pervade the entirety of our system.

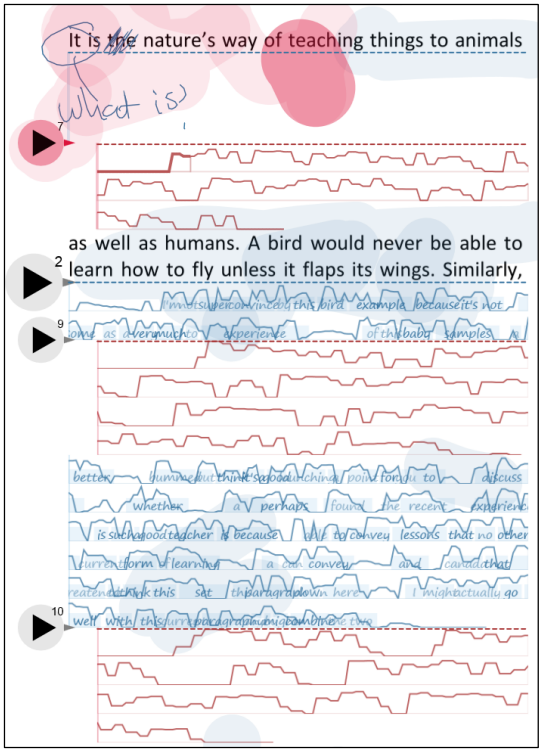


Figure 13. New annotations can be inserted under existing expansion space or in the middle of existing waveform (footage from the real-user data, P2); here the red user has replied to the blue user's existing voice comment.

4.4 Embracing fluid document layout[7]

To enable semantic access to speech, RichReview presents an audio stream as a waveform. However, this visually rich representation requires a large screen real-

[7] The text and figures of this section were derived from a previous publication (Yoon et al., 2013).

estate. When an annotation is anchored to a part of the body text, this interface is prone to occluding the surrounding texts and to prevent comprehending the context around the comment. One walk-around is to toggle the interface between a compact anchored icon and a pop-up navigation interface (Adobe, 2016a), but then the visual access is not readily available anymore, because this approach hides the navigation interfaces behind the icons. In other words, the rich visual interface is in a dilemma between fast access and anchored context. To breakthrough this problem, we present a fluid document layout technique called TextTearing that can interleave visual entities in the flow of texts.

When there is insufficient space for displaying annotations, TextTearing interaction (Yoon et al., 2013) can be used to create additional writing space in between lines of text. To accomplish this, we let users create an expandable region of whitespace by adjusting the spacing between lines of text. As the region grows, the content below it is shifted lower in the page (see Figure 14).

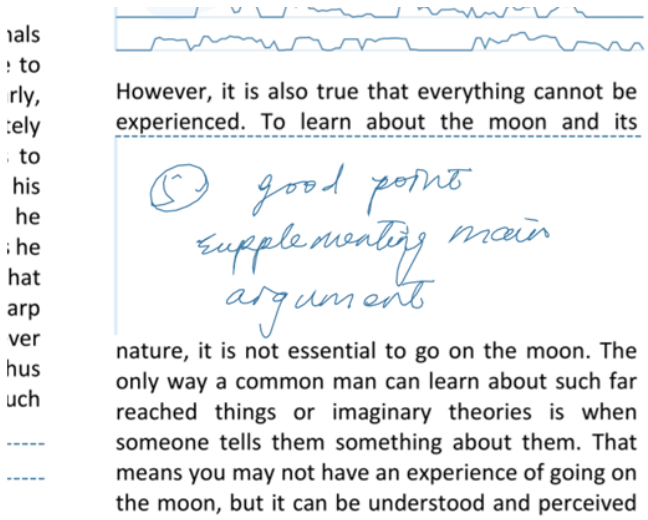


Figure 14. TextTearing space created in between lines of text.

Users can execute TextTearing interaction by drawing a horizontal line at the approximate place, followed by a pigtail in the vertical direction. All annotations (including the multimodal ones described later) are anchored to the nearest line of text, graphic, or expansion region on the page. When the position of these elements shift in response to re-layout, the anchored annotation also moves. Annotations created in this way can also be collapsed so that the original layout of the document is preserved.

4.5 Preliminary evaluation of RichReview prototype

We wished to determine whether users could successfully employ the features of RichReview to make comments on a document. Moreover, we wanted to investigate how these features were actually used. Therefore, we conducted a qualitative, formative user study using our prototype system.

4.5.1 STUDY DESIGN

To prompt realistic feedback from participants, we designed a task based on a representative classroom situation. We asked the participants to assume that they were working as a teaching assistant (TA) for an introductory, undergraduate writing class, commenting on a student's essay assignment.

Operating under the assumption that the social relationship between collaborators may influence annotation behavior, we told half of our participants that their annotations would be shared with the instructor of the course. We told the other half that their comments would be shared with a peer grader. The point of splitting our participants into two groups was to expand our coverage of possible usage scenarios rather than to carry out a controlled comparison or test a hypothesis.

Procedures

Our study consisted of a practice session that helped participants familiarize themselves with RichReview interactions followed by an open-ended session with full-fledged tasks. During the practice session, we introduced each feature of the system to each participant by demonstrating a specific use case, and then letting them try out the features first hand. For example, the Spotlight feature was introduced in the context of referring to a location of document while recording. Next, we gave each participant a set of practice tasks to carry out. We ended the practice session by having participants use the audio editing functionality: participants were asked to pick one of the most problematic recordings amongst the ones they had created, and edit it so that could be better understood. The practice session took less than 40 minutes.

The open-ended annotation session consisted of two parts. Participants started with the production part where they gave constructive feedback on an essay for 15 minutes. They were asked to create at least 6 comments on an essay concerning any kind of writing issue. Then, participants performed the consumption part, which looked at how rich annotations are consumed and discussed. We asked participants to listen to a set of pre-made annotations, and then respond with constructive feedback for 20 minutes. In both tasks, participants were not required to use any specific interaction techniques. We concluded the evaluation with a 10 minute of semi-structured interview session. In total, the study tasks required approximately 90 minutes to complete.

Materials

The materials used in the study were sample essays to the "Analyze an Issue" portion of the Graduate Record Exam used in the United States for entrance to graduate school. These essays tend to be around 500 to 650 words long (approximately 1 page) and of moderate writing quality. We picked two different essays, counter balanced across participants to rule out text dependencies. Each participant used the same essay across the production and consumption tasks, in order to save reading time.

The pre-made annotations in the consumption task were composed of various discussion topics and consisted of diverse modality combinations. These were based on real annotation data captured in earlier pilot tests of our system.

Participants

We recruited 12 participants from student mailing lists at Cornell University. The average age of our participants was 21.3 years old. All but one participant was a native English speaker and all had experience with collaborative writing tools. The most frequent discussion channel for their writing tasks was e-mail (5.33 hours/week), followed by F2F meeting (3.67 hours/week). The participants received $15 for taking part in the study.

4.5.2 RESULTS

Broadly speaking, the results of our study demonstrated that participants could successfully employ RichReview to communicate complex ideas about a document. Voice and Spotlight introduced additional expressiveness and efficiency on top of the

communication capability of the legacy collaboration tools that were based on textual means. Moreover, cross-modal commentaries and indexing features helped users achieve fluid modality combination and lightweight annotation access.

Experience of Using RichReview

Users compared the overall experience of using RichReview to having a collaborator virtually present. P3, when talking about the Spotlight feature remarked, "It (Spotlight) was like I was talking to someone in person when I point to an area." Further evidence of this was the fact that in annotations, participants often used the pronoun "you" reinforcing the sense that they were talking through the computer rather than talking to the computer.

Annotation Production

Ink Annotations. Direct inking without additional forms of recording was the most widely used form of annotation. All of the participants used ink for simple mark-up such as circling, underlining, question marks, brackets, proofreading symbols, connecting lines, and personal notes that they wanted to revisit. This highlights the importance of static inking for lightweight interaction.

Voice Annotations. Voice recording was used by all participants when they wanted to make a comment that was longer or more detailed. Participants praised voice's speed and expressiveness. As P4 said, "Now I can hear someone's voice and understand completely what they're trying to say versus just seeing their note and trying to interpret." All participants except P1 and P8 used voice in conjunction with writing and the Spotlight. These two users used voice on its own.

Participants structured their voice annotation in many different ways. One way was to write ink as a visual guidance for the verbal description. For example, participants first made underlines on body-texts or wrote down key points in white space while reading, and then used voice or Spotlight to refer to these points while recording. Another way was to write keywords during the recording session to allow their hypothetical recipient navigates to a certain topic in the recording by tapping on a corresponding keyword. Additionally, P3, P5, P8, P9, P11 used the feature where additional annotations could be appended onto an existing one; they used this feature when they had multiple points to talk about in a single paragraph.

Another interesting observation was that participants tended to use voice when they disagreed with an idea and ink when they agreed with it. The reason for this was because when they disagreed, they would use voice to provide a detailed explanation for their disagreement (P6, P7, P10, and P11). This suggests that support for voice annotations could be the best method for achieving group maintenance goals (Birnholtz et al., 2013).

Spotlight Annotations. Participants frequently used Spotlight when speaking. The Spotlight feature was used to refer not just to the underlying text, but also to other people's ink marks and a part of waveform or transcripts. Annotations about paragraph structure or logical inconsistencies were often accompanied by the spatial cues that Spotlight conveyed. Overall, the feature was seen to be a powerful deictic tool: As P3 said, "I liked that feature (Spotlight) a lot, because it could direct somebody while recording to the specific spot that they are talking about."

Participants did raise some implementation issues, however. P1, P9, and P10 reported that Spotlight was sometimes recorded inadvertently when the pen hovered over the screen for other reasons. P1 complained that the Spotlight trail was too thick to point to a specific line of text or word. In future iterations of the system, these issues could be addressed by filtering out spurious hovering gestures and by changing the blob size.

Socially-Driven Modality Choices. Besides the annotation content and purpose, we found that the social factors affected which communication modalities participants felt comfortable employing. For example, P2 and P8 regarded simple scribbles, such as circling or checkmarks, as an impolite or casual form of annotation, choosing instead to leave voice comments. By contrast, P1 and P8 thought that writing a complete message was politer than voice. In this case, they claimed written comments were easy to understand and that voice is a "lazy form (P8)". While we cannot draw any conclusions on this basis, this does raise some interesting research questions for future research on annotation and target users.

Editing Audio. Most participants found the voice editing interface easy to use and efficient especially for removing long pauses or utterances such as "Um". This suggests that automatic detection and trimming of the pauses might be useful. However, considering that some users depend on long pauses as a navigation cue for time indexing operations, removing these also might be problematic. Ultimately the long term usefulness of these features would need to be assessed in real practice. Later in the chapter 7, we present a follow-up study that tests real-world effectiveness of our novel voice editing interface.

Annotations Use and Organization: Annotation Positioning

Most annotations were placed immediately under the relevant text. If it was about a sentence or keyword, participants would position the annotation under a line in the middle of a paragraph. However, P1 and P8 were reluctant to break the paragraph structure, instead placing the recording below the paragraph and making a reference to the targets using inking or the Spotlight. Some annotations do not have an obvious anchor point such as when they are about multiple paragraphs or global writing issues. In these cases, participants usually placed the recordings below the end of the right column, making them hard to distinguish from those relating to the last paragraph. This observation suggests that a distinct space to anchor meta-commentary (Qixing Zheng et al., 2006), possibly close to the bottom of the page, might be useful.

Consuming Annotations

Navigation via the visual representation of the audio was actively used to jump into or revisit a voice annotation. Participants were able to use the waveform as a navigation cue effectively when there were salient features they could focus on, such stretches of silence. Other participants sometimes used the words in the transcript as a way to navigate (P7, P9, and P12).

However, most of participants preferred using the waveform over transcription because of the detrimental effect of transcription errors. For instance, P11 recounted one instance where the phrase “kind of this” was recognized as “Kennedy.” Although P11 was aware it was a transcription error, the participant found it very hard to ignore.

Participants found Spotlight trails useful for getting a sense of what an annotation was about. However, participants did not use ink or Spotlight trails to index into annotations. We believe there were two reasons for this: First, Spotlight traces became too cluttered; second, users were not familiar enough with the style of annotation to know how the ink or Spotlight element was structured in relation to the audio.

Creating Responses and Discussion Threads

The ability to create rich annotations about existing ones was well adopted by all participants. In general, most responses tended to be placed immediately below the annotation to which it responded. For instance, P2 inserted a voice annotation in the middle of an existing audio stream (Figure 13). P10 made a written reply below a part of existing spoken annotation making use of cross-modal commentary features (Figure 12). They used cross-modal mark-up features for referring to parts of spoken or written annotations. For instance, P7 used the Spotlight to point to the visual representation of the audio (Figure 11), and P10 marked it up with ink (Figure 12).

4.6 Summary and implications

In this chapter, we presented design of RichReview, a multimodal annotation system. RichReview allows users to mix flexible combinations of multiple interaction modalities including speech, inking, and gesture. Consuming the multimodal annotation is direct, simple, and fast thanks to a semantic and cross-modal indexing technique that uses waveform or transcripts as a surrogate for access to recorded voice. Employing fluid layout techniques creates document spaces for commenting,

which eventually resolves the limited screen real estate problem and also displays threaded conversations in the lines of texts and in the flow of the text.

A subsequent preliminary evaluation confirmed that users can express complex and nuanced ideas through RichReview annotations. Also, it was confirmed that the indexing feature balanced production and consumption so that users can access the voice data faster and easier. However, the lab study could not tell if the system truly benefits people by advancing their everyday jobs, because the tasks in the lab studies were dislocated from the context of real work process. In the next chapter, we go beyond the limitation of the lab studies by conducting a series of deployment studies.

5 Field deployment studies[8]

Although a small-scale formative laboratory study validated the interface concept behind RichReview (Randles et al., 2015; Yoon et al., 2016), the real-world efficacy and implications of the multimodal features it introduced (e.g., pointing gesture or audio visualization) have not been fully characterized. One great way to evaluate a system for real-world tasks is through field deployment. In the course of people adapting the new technology to fit it into their existing work process, we can understand what breaks down and what more is needed, as well as what benefits users really care about. Two representative classroom activities were selected as the target tasks: instructor feedback and peer discussion. These document-centered activities fit well with the use cases of RichReview as an annotation system. Moreover, the impact of RichReview can easily be compared to traditional methods since users can compare their experiences from the deployment studies with their previous experiences without the new system.

Goal of the deployments

This series of evaluations provides answers to the following questions:

• What practical benefits does the multimodal integration of voice and ink with gesture offer in realistic settings and for typical classroom activities (e.g., paper revision or peer discussion)?

---

[8] The text of this chapter was derived from a previous publication (Yoon et al., 2016).

• To what extent can waveform indexing overcome the problem of diminished accessibility in multimodal annotations? Are there any caveats?

• What are the broader implications of these features for the use of multimodal annotation tools for tasks in online classrooms? For example, what are the expected limitations and possible workarounds?

5.1 Deployment for instructor feedback to writing assignments

Our first deployment examined the utility of the new multimodal annotation features in the context of an instructor providing feedback on a term paper. The main focus in this study was students' perceptions of the effectiveness of the feedback, but we also observed the emotional responses and interpersonal dynamics that emerged from the use of the tool.

5.1.1 WEB-BASED VIEWER FOR RICHREVIEW

In the original RichReview system presented in the chapter 4, making full use of the capabilities of the system required specialized hardware (Windows tablets). While it would have been possible to provide each student in a small class with such a device, it was an inherently impractical and non-scalable solution considering deployment costs.

Our solution was to create a new, web-based viewer for documents annotated using RichReview. The instructor used a tablet to create her comments (leveraging the full feature set of the system) and uploaded the annotated documents to the web.

Students could then access instructor comments on their device of choice by visiting a URL the instructor provided.

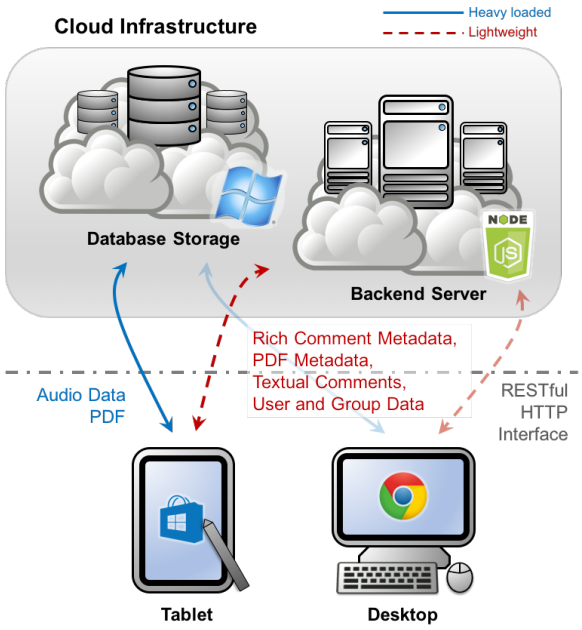


Figure 15. The cloud infrastructure of RichReview.net. Low latency data sharing is made possible by separating heavy loaded (blue) and lightweight (red) data channels. Additionally, the system supports web standards for cross-platform access, secured connection, and accessibility features.

Smooth and agile data sharing is essential for computer-supported cooperative work (CSCW) systems to support fluid collaboration. Exporting and sharing a set of multimedia data is a barrier for our users to share rich annotations. Inspired by the architecture of the United Slate system (Chen et al., 2012), we built a cloud infrastructure that handles the heavy loaded multimedia data and lightweight metadata in two separate channels (Figure 15). The centralized metadata storage enables prompt data sharing among the collaborators via weblinks, and the multimedia data are shared through the distributed database storage without causing network overload. The backend server was built on a node.js framework running on Microsoft Azure. The

cloud data storage was implemented using Microsoft Azure binary large object (BLOB) storage.

Our implementation of RichReview.net satisfied the security and accessibility requirements of the Family Educational Rights and Privacy Act (FERPA) in the United States. Every entity of the interface is a document object model (DOM) that is compatible with screen reader software. For secure access to student data, all authentication and data exchange procedures follow OAuth 2.0 and HTTPS protocols.

5.1.2 DEPLOYMENT PROCEDURES

The system described above was deployed in an undergraduate level Human Development course in the fall 2014 semester at Cornell. The term-paper assignment involved writing a proposal for a life-span developmental research project. The paper was to include prior literature, research questions, measures, methods, and plans for analysis. Recommended paper length was under 20 pages, and students worked individually. Students in the class submitted paper drafts as PDF files over email. The instructor then commented on each paper with RichReview on a tablet using digital ink, audio, and gestures. Students received feedback 10 days before the final version of the paper was due.

The instructor, who was a member of our research team, spent an average of 20 minutes commenting on each paper, but she noted that weaker papers appeared to take relatively more time than the others. She took advantage of the full range of affordances, often combining multiple modalities (e.g., drawing a flow chart and ‘walking’ students through the chart by speaking and pointing over the drawing).

Also, the instructor attempted to communicate emotion by expressing enthusiasm or encouragement. The absence of a voice editing feature required the instructor to start recordings from scratch if she felt the comment was unclear or got side-tracked. Although she generally regarded this as a shortcoming of the system, she noted that the re-recorded comments were often more cohesive and of better quality.

5.1.3 PARTICIPANTS

Participation was voluntary and students could choose between receiving handwritten notes on a printout or RichReview-based comments. Exit surveys were also optional and the instructor did not have access to student responses. The class was composed of one instructor and 17 undergraduate students, 16 of which participated in the study and 13 of which answered the exit survey. Among the participants, there were 12 females (mean age: 21.0, SD = .44); one student was male, and one was a graduate student. All students used laptops to view the feedback provided.

5.1.4 MEASURES

We monitored and logged students' online activities such as page navigation and voice playback. The exit survey collected subjective ratings on a 5-point Likert Scale. Students were also asked to rate their preference for the different ways of receiving feedback from the instructor. Since students had received paper-based handwritten feedback from the same instructor earlier during the semester, they had a baseline against which to compare RichReview-based feedback. Lastly, the questionnaire also featured free-response questions about the usability of various system features (See the auxiliary material for the copy of the questionnaire).

5.1.5 RESULT

Essays ranged from 10 to 22 pages in length (M = 15.7 pages, SD = 2.73). The instructor made digital ink markups for typographical edits as well as voice comments for detailed commentaries. On average, the instructor made 51.4 voice comments (SD = 11.7) per student essay with a mean length of 14.9 sec (SD = 15.1).

Students reported that the system was easy to learn and effective. One commented that "This was by far the best experience I've had while revising a paper (P10)". Students reported a willingness toward continued use and recommended its use to peers. To quote P6, "Would definitely use again and would recommend this to others!"

Benefits of Multimodal Indexing for Consuming Comments

One recurring theme in the qualitative feedback we gathered was that easy-to-use audio replay was helpful for consuming the recorded comments. We analyzed logs of online activities to take a deeper look into replay patterns. The results showed that audio re-listening was very popular. Most (73.1%) of the voice comments were replayed more than twice, and a few (6.44%) were replayed more than 10 times. As shown in the replay count histogram of Figure 16, more than half of the users (9) played the majority of the comments they received more than twice. Moreover, the histogram for every student featured a few comments that were replayed many times. P5, for instance, predominantly listened to comments once, but had two comments that were replayed more than 14 times.

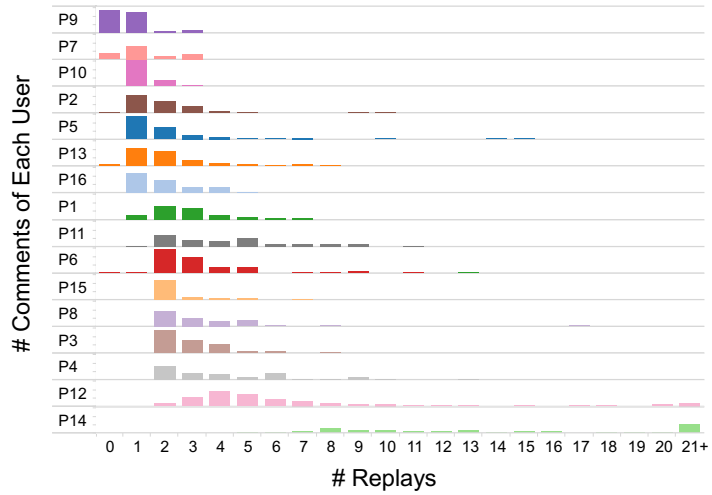


Figure 16. Each row indicates a distribution of the number of comments (y-axis, ranges from 0 to 32, un-normalized) per the number of replays for each user (x-axis, cut off at 21). The rows were sorted in the order of momentum position.

Analysis of data revealed that the students exploited various types of indexing features for re-listening, but made particularly heavy use of audio waveforms. Across all playback operations, the play button was used 81.7% of the time, waveforms were used 12.4% of the time, and Spotlight traces were used 4.6% of the time. We also found that the number of replays started by clicking on the waveform for a given voice comment was significantly correlated with the comment's length (N = 823, $r = 0.22$, $p < .001$). A similar pattern did not emerge for the play button. This suggests that for long comments, students perceived waveform-based indexing as useful in jumping straight to relevant passages in the audio.

To further illustrate this point, we show a representative user's (P14) pattern of accessing a set of annotations (Figure 17). With a lengthy comment like comment #1 (58.5 sec), P14 jumped into a specific point multiple times using Spotlight indexing (3.6 sec, 7 times, the orange bar) and Waveform indexing (13.4 sec, 16 times, the red

bars). The clusters of bars with both green and red in comments #2 and #3 show that P14 triggered playback using a mix of controls. The tall green bars on the left side of Figure 17 indicate that the comments are almost exclusively played first using the playback button. Finally, the green bars in the middle of recording show that P14 employed a stuttering playback pattern in which P14 repeatedly paused and started the comments in the middle of the audio stream. Similar patterns were observed in the vast majority of the other users (P1, 2, 4, 5, 6, 11, 12, and 13).

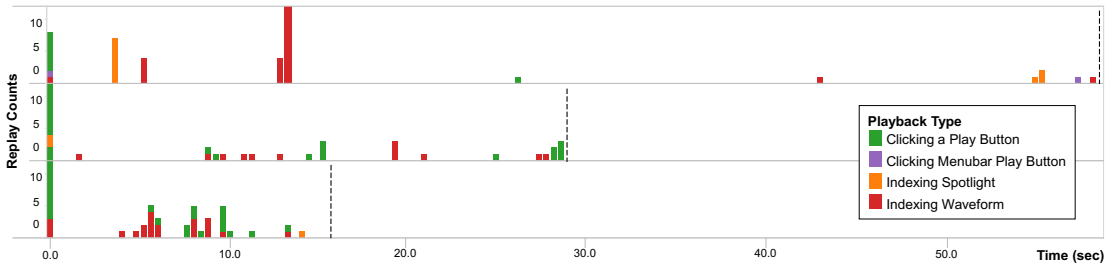


Figure 17. Navigation patterns observed from a user listening three different recorded multimodal annotations. The x-axis is time separated into 0.4 sec interval (ranges from 0 to 58.2 sec), and the y axis is a number of playback hit in each time interval.

These quantitative findings were further corroborated during our survey. Participants reported that they used the waveforms to "skip over the parts you adjusted already (P6)", "listen to specific parts over again (P3)", or "repeat a missed word (P4)". Participants ratings also suggest that the Waveform indexing feature was very helpful for understanding audio comments (M = 4.63, SD = .52, 8 responses). On the other hand, the gesture-based indexing using Spotlight traces was not as popular as the waveform indexing. Only 9 participants noticed the feature and used it. Participants reported that the feature's lack of discoverability was the barrier: "I didn't know you could do that (P8)".

Multimodal Annotations vs. Longhand Comments

In the exit survey, we asked participants to compare their preferences between RichReview and traditional feedback methods for different types of comments. They preferred the multimodal comments over written comments for receiving feedback about writing issues related to factual content (M = 4.27, SD = .79) and structure (M = 4.55, SD = .52). There was no significant preference difference between the two methods for comments pertaining to grammatical errors and typos (M = 3.09, SD = 1.45). This result echoes and confirms previous research on voice-only annotation (Chalfonte et al., 1991; Kraut et al., 1992; Neuwirth et al., 1994), which found that spoken comments were preferred over text when describing higher-level (structural, or semantic) writing issues in comparison with local problems.

Multimodal Annotations vs. Office Hours

Surprisingly, the majority (11) of participants preferred RichReview annotated documents over office-hour meetings (M = 3.91, SD = 1.22), and believed that they offered an acceptable substitute for in-person meetings (M = 3.91, SD = 1.04). Qualitative comments offered two explanations for this result. First, students wanted to "incorporate all the comments" and make sure that they were "doing everything that the instructor suggested" (P4). RichReview was useful because recorded comments were hard to miss and could be addressed one at a time. Also, the rapid stream of feedback received when meeting in person made students worry about whether they were missing or misunderstanding the instructor's comments. In contrast, the recorded comments could be replayed multiple times if they were not clear. To quote P4, "I can

listen to everything multiple times if I didn't get it which also made it less intimidating". On the other hand, a few students (2 of 13) thought that in-person meetings offered a more immediate interactive dialogue which RichReview's asynchronous interaction did not offer.

5.2 Deployment for TA feedback on math assignments

The previous deployment promised the benefits of the system for the essay feedback process in a small social science class. To test if these findings can be transferred and generalized to another context, a follow-up study in a different setting was required. This chapter presents a follow-up deployment of RichReview for an assignment feedback process in a large math class.

Through a campus-wide recruitment for a class willing to use RichReview as a feedback tool, we met a math professor who found the rich communication capacity of multimodal commenting to be a viable approximation of what she does in face-to-face instruction for assignments and prelim grading in her introductory vector calculus class. For example, she *draws* graphs or *doodles* equations while *pointing* at them with a finger to *speak* about complex math concepts. Her insight was bolstered by the literature in educational psychology, as researchers found that using the combination of speech and gesture leads to enhanced learning in math instruction (Cook et al., 2008; Goldin-Meadow, 2005).

The introductory vector calculus class was particularly suitable for our deployment thanks to its large size. The target class has regular enrollments of more than 100 students per semester because it is a required course for all engineering students at Cornell. The large sample size allowed us to measure the impacts of the different feedback tools with statistical significance, augmenting or going beyond the previous findings of the small-scale deployment. The measures were geared toward concrete

indicators for efficacy of the tools including students' perceived efficacy, learning gain, and instructor–student relationships.

As the classroom context moved to the large math class, we adapted the experimental tasks to the new deployment setting. The research team had two one-hour meetings with the instruction team—the professor and teaching assistants (TAs)—to understand the current feedback process of the class, close the gap between the differing goals of researcher and instructor, and create action plans. All these lessons helped us make informed decisions.

We learned that students were doing homework in notebooks, putting them in a submission box in the classroom, and getting it back from the assignment returning room. The instructors were giving longhand feedback directly on hardcopy submissions. Hence, similar to the previous deployment study, we decided to compare students' perceptions of the rich multimodal feedback vs. the traditional pen-and-paper feedback. Showing the results from the previous deployment motivated the instruction team to actively participate in the study, as the professor thought that the communication modalities of RichReview would be optimal for delivering complex math concepts taught in her class.

We decided to run a semester-long deployment to collect enough data. When deploying the tool to a new class for the long term, it is imperative to give the instruction team and the stakeholders of the course the conviction that it does no harm to student learning. We thus decided to conduct a short dry-run deployment to check if our experimental practice, including use of the RichReview feedback, caused any critical problems. To minimize the risk of using the experimental tool for real

classroom activities, the preliminary study deployed RichReview only for one week-long assignment feedback process, while the primary study deployed RichReview for ten weeks involving both assignment and prelim feedback processes.

5.2.1 BUILDING A NEW PAGE LAYOUT ANALYSIS MODULE

Unlike the previous deployment setting where the students' essay submissions were PDF files generated from word processing software, the submissions in this study were unstructured page scans of the hardcopy papers. The problem was that the student-generated PDF files lacked the page layout metadata, which is necessary for inline commenting of RichReview. This led us to design and build a new document layout analysis module to detect text lines of assignment pages.

Detecting page structure of the math submissions could have been daunting, because text lines of students' freewriting were ambiguous and ill-defined. Fortunately, most math students were doing their homework on lined notepads, which provide a clear visual structure of the horizontal strips. The level lines have many unique features that make them visually distinctive; the set of straight lines are parallel and evenly dispersed. Using the lines as strong morphological constraints, we built a page layout analysis module that converts a scanned PDF into a RichReview document. For a given page image, the module first detects all possible straight lines using the Hough transform (Duda & Hart, 1972). Then it applies RANSAC on the Hough space to extract the notebook lines with the same angle and regular intervals (Fischler & Bolles, 1981). The parameters of algorithms were heuristically adjusted based on the page scans of the student assignments from the past semester.

5.2.2 PRELIMINARY 1 WEEK DEPLOYMENT

For this preliminary study on the fall 2015 semester, we recruited 14 student volunteers in the calculus class. The TAs generated RichReview feedback for student homework submissions by creating multimodal annotations on the digitized assignment documents. The procedure is as follows:

(a) After finishing the homework, every participating student scans her/his assignment document into a PDF file and submits it to the RichReview course management website.

(b) The instructors (TAs) generate RichReview comments during the grading process. After the grading is done, they upload the comments to the RichReview website and announce to the students that the feedback is out.

(c) The participating students sign in to the website to review/listen to the recorded feedback. Students can access the comments as often and as long as they like.

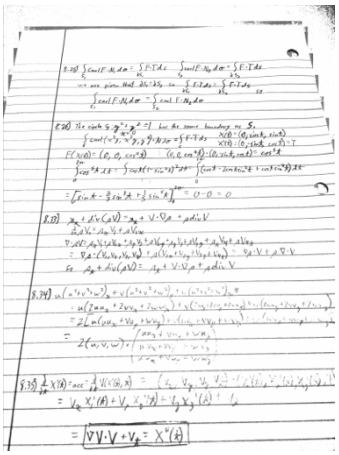
For click stream level data collection, our web app logged students' online activities while they are reviewing the feedback. We also asked the students to rate perceived quality of the feedback experience through an online survey sent after the week-long deployment.

Results

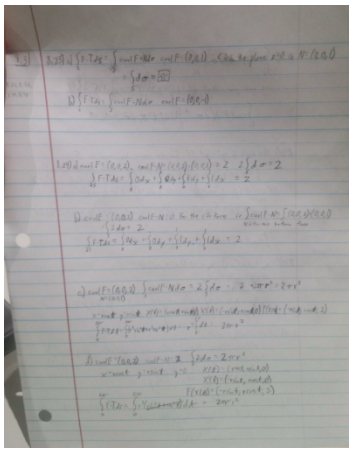
Fourteen undergraduates participated to the study (M_{age} = 19.27, SD = 1.10). For each assignment, the TA made 1.43 comments on average (SD = .51) that lasted 12.13 seconds (SD = 5.99) and had 1.45 gesture strokes (SD = 1.15) on average.

The survey responses indicated that RichReview can be an effective alternative to pen-and-paper feedback. Eleven survey responses were collected; the respondents found voice comment significantly easier to understand than written comment ($M_{RichReview}$ = 4.3, SD = .5 vs. $M_{pen\text{-}and\text{-}paper}$ = 3.4, SD = .88, Wilcoxon Z = 3.0, p = .019, Cohen's d = 1.2). Qualitative responses echoed the benefits of RichReview found from the previous study, mentioning that voice was easier to understand (2 of 11 respondents) than handwriting which is often messy and illegible (6). Being able to repeat the voice was useful for reviewing the comments (2), and the pointing feature was also helpful (2). It also felt more "personal [P2]" and "feel[s] like the instructor was in the room with you giving you direct feedback about your assignment [P4]," indicating the delivered presence of the commentator.

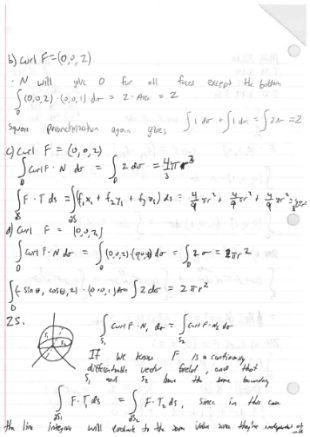
The positive student rating for RichReview feedback gave the instruction team a "go" signal for the semester-long primary deployment in the subsequent semester. The professor promised active cooperation for recruiting students for the study, and updated the course syllabus to publicize the use of RichReview feedback.



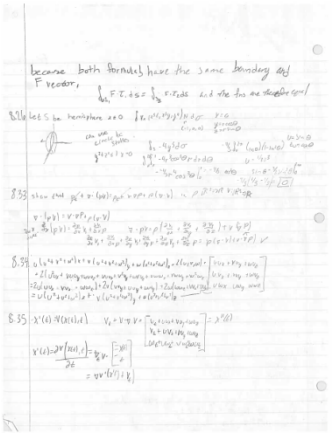
(a) slanted



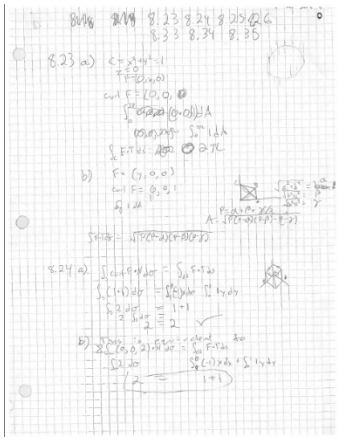
(b) blurred



(c) no line



(d) faint lines



(e) too many lines

Figure 18. Types of abnormal scanning results that failed our document layout recognition engine.

However, the PDF scanning/submission process needed fixing because using the mobile scanning app was cumbersome for the students while image quality was not good enough for reliable layout detection. Our page layout recognition app had failed to detect 5 out of 19 submission PDFs under several abnormal scanning incidents as shown in Figure 18. From the survey, 5 of 11 respondents reported frustration with the mobile scanning process (e.g., “scanning and uploading process was annoying [P7]”).

We therefore designed a semi-automatic digitization process for reliable and scalable digitization of the submission documents.

5.2.3 IMPROVING THE ASSIGNMENT SUBMISSION PROCEDURE

For reliable page scanning in the semester-long deployment, we focused on solving the representative issues shown in Figure 18. To resolve the cases of low quality images such as (a) and (b), we decided to collect the assignment submissions in hardcopy, so that we—not the students—can digitize the process using a high-quality professional scanning machine. This process can be done semi-automatically by streaming the stack of assignments in the document feeder. To solve the problem of unusual notepads as in (c), (d), and (e), we asked the course instructor to guide students, with clear guidelines and announcements, to a type of notepad with uniform and easy-to-scan lines.

5.2.4 PROCEDURES FOR A SEMESTER-LONG DEPLOYMENT

The semester-long (nine weeks) main deployment targeted the subsequent opening of the same course in the spring 2016 term. Participants were separated into two experimental groups (A, B) for a comparison of the feedback methods using multimodal annotation on RichReview vs. longhand comments on hardcopy. There were 105 students enrolled in the course. The instruction team included a professor and two graduate TAs. The professor worked primarily on the lectures, and TAs evenly divided the grading responsibilities for assignments and prelim grading.

Semester schedule

When designing the experimental conditions, we aimed to provide a fair learning opportunity to students in both groups. A semester-long treatment can unevenly influence student performance, possibly in favor of the students in the RichReview group, considering the promising results from our previous deployment studies. As a solution, we split the semester into two four-week blocks to permute the order of the between-group treatments as shown in Table 1. Alternating the conditions balances out potentially competing effects of the two feedback methods by allowing students to experience both methods.

Table 1. The semester schedule of the deployment study.

weeks	5th	6th	7th	8th	9th	10th			11th	12th	13th	14th	
groups	Prelim1 (Corr.) 02/25	hw6 03/04	hw7 03/11	hw8 03/18	hw9 03/25	Prelim2 (Score) 04/05	hw10 04/08	Prelim (Corr.) 04/05	hw11 04/15	hw12 04/22	hw13 04/29	~~hw14 05/06~~	Final (Score) 05/16~24
A	RichReview						Pen + Paper	Pen + Paper					
B	Pen + Paper							RichReview					

The deployment began at week five of the semester, right after the first prelim when class enrollment becomes stable. The rest of the semester was separated into two blocks: the 1st block between the 1st and 2nd prelims, and the 2nd block between the 2nd prelim and the final. Each block contains four weekly homework assignments (HW) and an exam (a prelim or the final). Hence, the efficacy of the homework feedback scheme for the four weeks can be measured as the score of the following exams: 2nd prelim for HW6~9 and final for HW11~13. Before the beginning of the 2nd block, we had a recess week for HW10 where the entire class got pen-and-paper feedback while

the group B students were instructed about the protocols of the forthcoming RichReview deployment. The number of HW ended up unbalanced in favor of group A as the last homework (HW14) was unexpectedly canceled at the instructor's discretion to encourage students to focus on the final.

Tasks

The class gave instructor feedback for grading the two types of student submissions: weekly assignments (HW) and prelim. The weekly assignments let students review the subject matter covered in the past week's lecture by solving several exercises from the chapter of the textbook. The turnaround time for each assignment submission and feedback was one full week (Friday out and Friday in). For the prelim, the class used a standardized lined notebook called a Cornell blue book. The TA grading for each problem included constructive feedback describing the reasons for deducting points.

The prelim-revision is a thoughtful correction process that takes places right after taking an exam. Students can earn back up to 50% of the deducted points by writing a report that diagnoses their errors and goes over the solution. It is worth noting that the RichReview was used for grading the prelims, not the prelim-revision reports, which was done in the traditional way throughout the semester. The sum of the homework grades was worth 30%, two prelims were 40%, and the final was 30% of the total grade.

Measures

The dataset collected from the deployment includes scans of the student submissions, rich comments from the TAs, students' activity logs, their responses to post-deployment surveys, and scores they earned from assignments, exams, and prelim corrections. Specifically, the data collection process was geared toward capturing several indicators for efficacy of different feedback tools. The survey questionnaires asked students about subjective ratings of their perceived efficacy using the two competing tools. The rating had ordinal factors from "Strongly Agree (5)" to "Strongly Disagree (1)." The survey questionnaires also included open-ended questions to gather qualitative implications of using the tools. The scores students earned from the assignments and exams were collected as a measure of learning. The prelim score could be a good indicator of a long-term effect (1~4 weeks) of assignment feedback, while the prelim correction score could be related to a short-term effect (1 week) of the feedback from prelim grading.

Recruiting participants

We recruited the participants from 105 students enrolled in the class; 74 students volunteered. They were randomly assigned to the two experimental groups, constituting a group of 37 for each condition. However, three students from group A and five from group B dropped the class before the end of the semester. We used the data only from the students who got grades by completing the course. In other words, the following analysis is based on the data from 34 group A and 32 group B students.

5.2.5 RESULTS

The students turned in most submissions. They never missed any prelim, perhaps because it contributes to a significant portion of their final grade (66 prelims submitted). Most homework submissions were also turned in (93.5%, 217 out of 232 assignments submitted). The new document digitization pipeline worked efficiently and effectively throughout the semester. Most submissions were scanned, structured, and converted to a RichReview document successfully without human intervention, but manual process was required to check before the release. Every week, it took about five hours to process the submissions of the entire class. The majority of the working hours were spent for the manual checking procedure. No student or TA reported a significant problem from the results of the digitization process.

Overall trends

Each instructor feedback had about two RichReview comments per submission ($M_{\#comment}$ = 1.96, SD = 1.85). Each of the RichReview comments had ~9 seconds of voice recording ($M_{duration}$ = 9.20, SD = 6.38) and 3.29 strokes of deictic gestures (SD = 4.47). Besides the rich comments, the instructors also used static ink writing for graphical mark-ups and scribbles in support of the rich descriptions ($M_{\#ink\text{-}strokes}$ = 57.38, SD = 50.07).

By listening to the recording, we could observe a trend that the contents of rich comments differ by length of the recordings. When a comment has a lengthy recording of multiple sentences, the speech usually delivered elaborated explanations on how to improve the student's answer or clarified misconceptions found from the answer. Most

of the multi-sentence comments accompanied gesture strokes referring to parts of a student's answer (see Figure 19). TAs also often made a brief audio recording to leave a short encouraging message, such as "Well done," or "Good job." Most comments shorter than four seconds were such notes without complex mathematical descriptions.

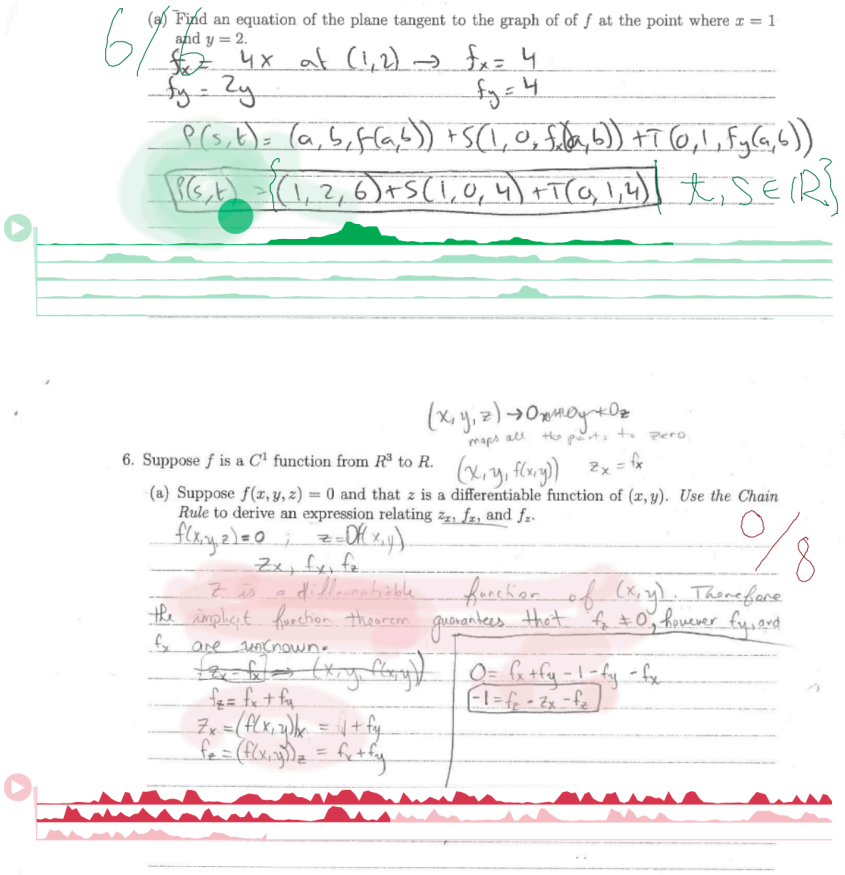


Figure 19. Example TA comments made on student prelim submissions. TAs exploited a combination of speech, gesture, and inking for describing how to improve the solution.

We conducted a comparative analysis for number and duration of TAs' rich comments for the two groups (A and B). To analyze commenting patterns between A and B conditions, we used unpaired *t*-test (two-tailed). When there is a significant difference, we also report effect size of the comparison measured using Cohen's *d*. To

control the experimentwise error rate, we used Bonferroni adjusted p-values for multiple hypothesis testing.

On average, group A's submissions received a significantly greater number of rich comments for the semester (M_A = 11.68, SD = 3.62 vs. M_B = 5.81, SD = 2.32, t = 7.79, $p < .001$, d = 1.92, Figure 20, left), a greater number of comments per submission (M_A = 2.34, SD = 0.72 *vs.* M_B = 1.45, SD = 0.58, t = 5.45, $p < .001$, d = 1.34), and longer total duration of recordings (M_A = 21.68, SD = 9.38 vs. M_B = 13.10, SD = 6.38, t = 4.32, $p < .001$, d = 1.06, Figure 20, right) than group B's.

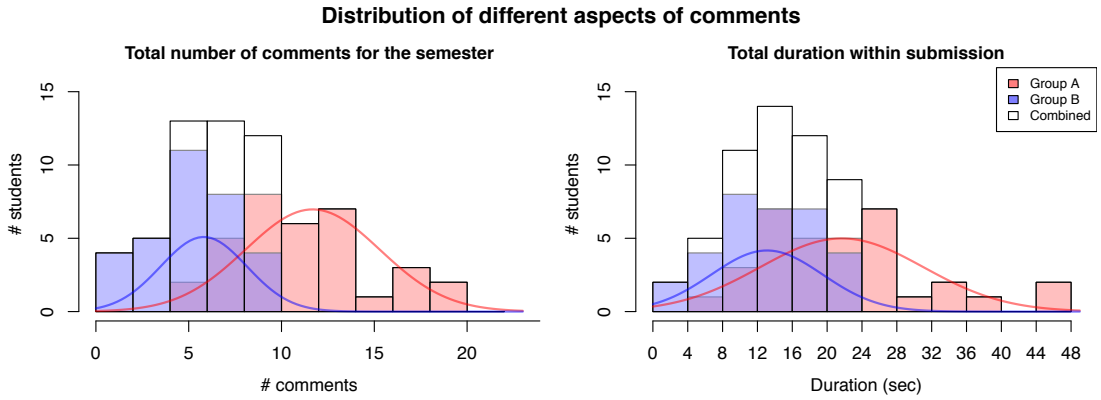


Figure 20. Histogram for the distribution of total number (left) and duration (right) of RichReview comments. The total number of comments is the sum of entire comments given to a student for the semester. The total duration is the sum of comments made on each submission averaged over the multiple submissions for the semester. Group A students (red) received more (left) and longer (right) comments.

This trend was mainly driven by differences in commenting patterns for the two prelims (see Figure 21, left, Exam 1 and Exam 2). The TAs reported that they could invest more time for grading Exam 1 than Exam 2 because they had more time for the TAing job in the earlier half of the semester before they get busy dealing with their own academic responsibilities as graduate students. Due to more comments in the first

prelim, and the canceled HW14, groups A ended up receiving about twice the number of total comments through the semester compared with group B (M_A = 11.68, SD = 3.62 vs. M_B = 5.81, SD = 2.32, t = 7.79, p < .001, d = 1.92).

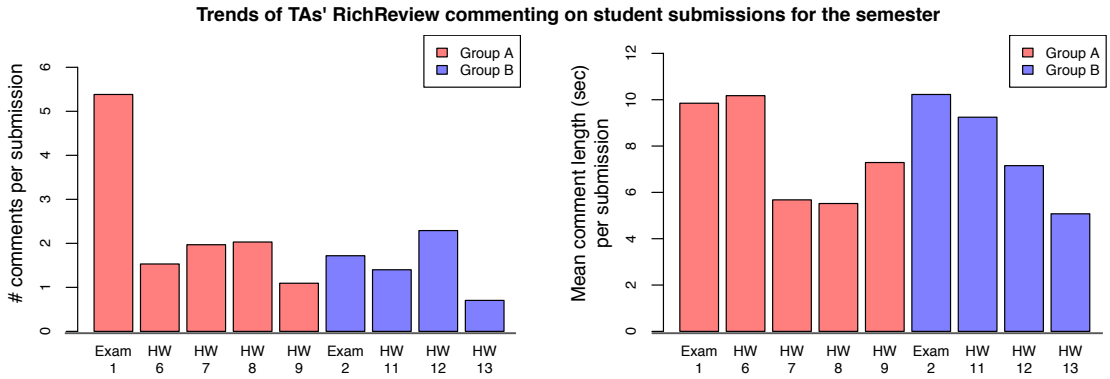


Figure 21. Trends of RichReview commenting for the deployment semester. The bar charts depict the number of comments per submission (left) and average comment length (left). The chart items were sorted in a chronological order from Exam1 to HW13.

Survey ratings

A total of 19 students in group A and 18 students in group B responded to the post-deployment survey. Based on this survey data, we tested if students in the different groups (A vs. B) perceived the two feedback tools differently.

Overall, the group A students rated RichReview higher than pen-and-paper. We first examined ratings for aspects of the two feedback tools within group A. The ratings from group A were analyzed using a two-way repeated-measures ANOVA with factors of tool (2 levels, RichReview and pen-and-paper) and questionnaire (3 levels from "Helpful for understanding," "Make me feel valued as a person," to "Convey emotion effectively," see x-axis of Figure 22). Mauchly's test indicated that the assumption of sphericity had been violated (p = .002), therefore degrees of

freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .78$). There were significant main effects of the tool ($F(1, 36) = 10.10$, $p = .003$, $\eta^2 = .10$), but the interaction effect ($p = .59$) was not significant, indicating that the ratings for RichReview were higher than pen-and-paper in all aspects of perceived efficacy in general. In other words, the students felt that RichReview, in comparison with the traditional feedback method, helped them better understand the TAs' points and emotions, and, further, gave them a sense of being valued as a person.

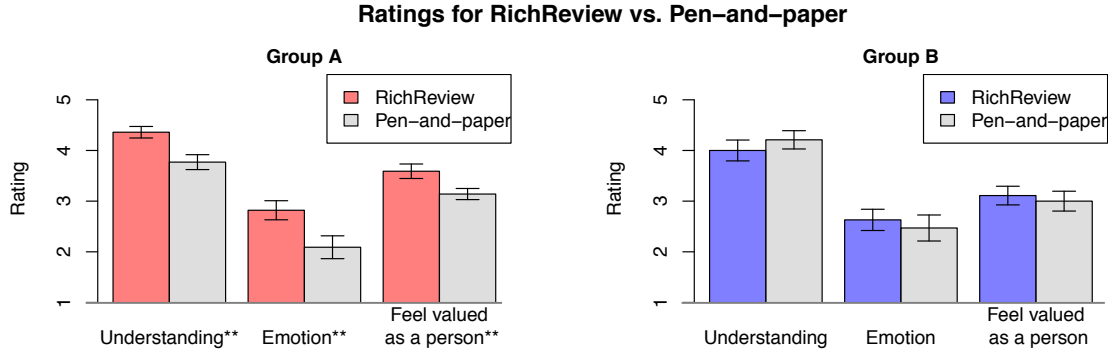


Figure 22. Ratings for ease of understanding for the helpfulness of RichReview for the coursework (between group comparison, left) and for different types of feedback tools (within group comparison, right). The error bars depict 95% confidence intervals.

On the other hand, the students in group B were missing such benefits of rich feedback. The group B students' ratings for the two tools did not show any significant difference. There were no significant effects of the tools ($p = .76$) or the interaction effects ($p = .59$). This disagreement between the two student groups could be attributed to the differences in their feedback experiences. From the analysis on the TAs' commenting pattern, we know that group A students earned more and longer RichReview comments than group B.

To further understand how the number and duration of rich comments can affect students' perceptions toward RichReview feedback, we examined the relationship between students' ratings and the aspects of the comments that the students received on their submissions.

As the first step, we drew a scatter plot (Figure 23) of each student's RichReview rating and mean recording duration from the set of TA comments made on their submissions. The ratings from the entire population (group A and B) were positively correlated with the total number of comments for the semester (Spearman's correlation, $r = .47$, $p = .005$) and total duration of comments per submission ($r = .41$, $p = .016$, Figure 23, note that each data point corresponds to each student).

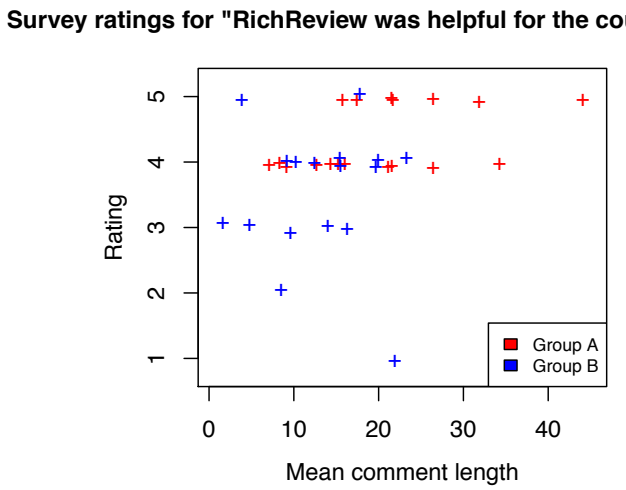


Figure 23. Scatter plots of ratings for RichReview from 35 survey respondents show positive correlations between the ratings and length of comments. The ratings for helpfulness of RichReview feedback were significantly correlated with mean duration. The data points were jittered to avoid over-plotting.

Yet this correlation might be led by the polarized responses from the two groups of students (i.e., higher overall ratings from Group A than B), where there might have

been other influential factors besides the number and duration of the comments, such as high academic pressure or harder subject matter during the end of the semester when Group B's data were collected. To rule out the biases from the discrepancy between the two group's experiences, we conducted within group analysis of the correlations. The Spearman test over the responses from group A showed that positive ratings (r = .49, p = .047) were significantly correlated with the total duration of comments (Figure 23, red data points), but the number of entire comments was not. Group B's rating was not related to any of the measures (Figure 23, blue), perhaps because most of them had too few and too short RichReview comments.

Exam scores

Between group comparisons of the exam scores did not show a trend that the feedback methods affected the students' scores. The prelim 1 score indicates the initial performance of each group prior to any treatment. The randomly assigned group populations were skewed in favor of group A, as shown in Figure 24 (left). The average prelim 1 score of group A was significantly higher than that of the group B (M_A = 68.4, SD = 17.8 *vs.* M_B = 58.1, SD = 20.0, p = .02, d = .54). After the four weeks of HW feedback (RichReview for A, pen-and-paper for B), the prelim 2 score did not show a significant difference (p = .17). After the second round of HW feedback (pen-and-paper for A, RichReview for B), final exam scores do not show significant difference (p = .33).

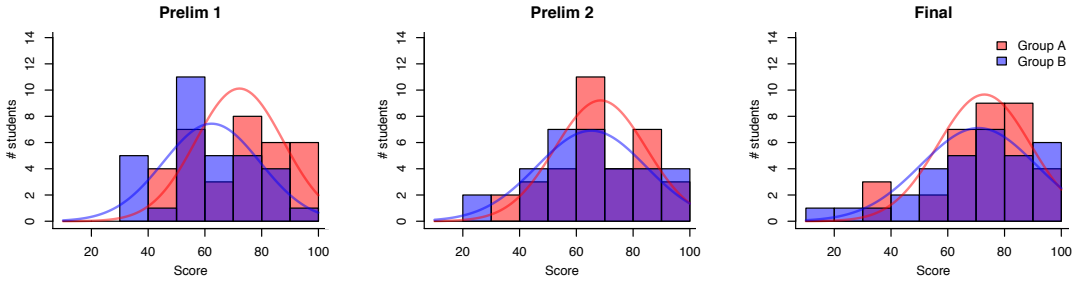


Figure 24. Histograms of the exam scores throughout the semester.

We expected that the feedback methods used for grading the prelims might affect the score students earned by submitting the prelim correction reports. However, the feedback tool was not the primary determinant of the correction score, because most students received an almost perfect score for the prelim correction reports. As they could earn 50% of deducted points back, the Δ score before and after the prelim-revision process had an almost linear relationship with the points they lost from the exam (linear regression, $R^2_{\text{prelim 1}} = .983$, $\beta_{\text{prelim 1}} = -.508$, $R^2_{\text{prelim 2}} = .949$, $\beta_{\text{prelim 2}} = -.494$, and $p < .001$).

5.2.6 DISCUSSION AND ANALYSIS

The results of the study confirmed that RichReview was best received by students when the instructor recorded longer explanations. This trend suggests that multimodal commenting works best for conveying complicated and nuanced concepts which take a long time to describe, echoing the benefits of multimodal annotation found from the previous essay feedback study (Yoon et al., 2016). This implication is also supported by the findings from the literature that voice comments works best for describing high-level concepts than for simple matters (e.g., semantic and structural matters of writing rather than type-writing errors of writing)(Chalfonte et al., 1991; Kraut et al., 1992).

Therefore, the use of rich feedback will be optimal for the setting where the subject matter of the feedback activity is complicated and demanding (e.g., research paper peer-review or senior-level/advanced courses), or when ample resources (e.g., time and availability) were available for the instructors to give students devoted guidance though rich feedback as in the earlier half of the semester of this deployment study.

Secondly, from the deployment of the new digitization process, we could learn that simplifying the submission process will greatly reduce the efforts of the instruction and research team. In our study, the manual digitization process was tedious and time consuming because we had to scan, as well as keep and manage, the hardcopy submissions. For efficient transactions of documents, a purely digital submission process is desirable (e.g., using LaTeX for math homework). If the target course demands hardcopy submission then using automatic scanning machines for digitizing a standardized notebook (e.g. Cornell blue book) will be necessary.

One limitation of this study is that we could not find effects of the feedback tools on learning gain. Future deployments can aim to measure the impact of the tools on students' learning gains if the feedback activities are relevant to the grading and have short turnaround time. In our deployment, the Δ score in prelim correction process was governed by the number of points they lost from the missing questions, not by the tools. Also, the feedback on the HWs were given to the students several weeks ahead of the prelim, which made it hard to measure the immediate impact of the feedback to the prelim score.

5.3 Summary

To test the real-world efficacy of RichReview, we deployed the system to the instructor feedback processes in two different classes. In the first study, an instructor gave rich feedback on student essays in a small social science class. To make this field deployment possible, we built a web-based viewer system for students to access the instructor's multimedia comments. The results suggested that students perceived the rich feedback as more helpful than the traditional pen-and-paper feedback thanks to the ease of access made possible by the visual indexing features of RichReview. Some of them even preferred it over face-to-face feedback, because asynchronous recording can be revisited anytime at their own pace. The second study was targeted to evaluate the efficacy of RichReview feedback, in comparison with pen-and-paper feedback, for students' prelim and assignment submissions in a mid-size math class. We built a page line recognition app for the digitization of hardcopy submissions, and a course management system for students to access to the rich feedback. Analysis of survey data and activity logs showed that the total duration of multimodal recording is indicative of students' positive ratings to the multimodal feedback. Overall, our findings inform guidelines for design and use of the multimodal commenting system for enhancing instructor feedback in educational settings.

6 Supporting online peer discussion with multimodal interactions

Through the two instructor feedback studies, we characterized how expressive communication modes of in-person meetings can be transferred into digital tools to provide rich and personal learning experiences for different types of classroom activities (e.g., giving elaborated feedback on students' essays or math assignments). In massive open online course (MOOC) settings, however, the lack of such learning experiences is the major deterrent to student engagement, motivation, and retention (Kizilcec, Piech, & Schneider, 2013; Zheng, Rosson, Shih, & Carroll, 2015). Hence, a natural extension of our work is to explore ways to offer rich and expressive pedagogical interactions to MOOC students. The challenge of extrapolating the setting of the previous study to a MOOC is that the instructor feedback is a one-to-many pedagogy that doesn't easily scale up to the sizes of MOOC classes due to a high student-to-teacher ratio. This chapter chronicles our endeavor to repurpose RichReview for peer discussion, which is another document-centered collaboration activity in classroom, but employs a symmetrical, and thus more scalable pedagogical model than instructor feedback.

As an alternative to the instructor's interventions, online classes offer personal and interactive experiences through community-based peer learning pedagogies where students grade each other's assignments and discuss shared interests (Stahl et al., 2006). In MOOCs, popular ways to structure the between-student interactions include synchronous online chats (Coetzee et al., 2015; Kulkarni et al., 2015), discussion

forums (Coetzee et al., 2014; Huang et al., 2014; Mak et al., 2010), and anchored discussion systems (Brush et al., 2002; Zyto et al., 2012). The asynchronous and in situ nature of the anchored discussion has promised many advantages, such as flexible scheduling, learning at one's own pace, and rich contexts of underlying academic texts. The great opportunity for RichReview is that most existing anchored discussion systems are purely textual, missing the rich interaction capacity that our annotation system can offer.

In this chapter, we present a redesign and a reimplementation of RichReview aimed at supporting rich online peer discussion at scale. To find user needs and potential challenges in the new task setting, a preliminary study evaluated a prototype multimodal discussion system against a weekly online discussion activity in a small social science class. We found that students prefer text over multimodal commenting because recording speech causes high speech anxiety and editing recorded speech demands more effort than text editing. We thus designed and built a novel re-synthesis-based speech commenting interface that delegates speech narration to a text-to-speech engine. A subsequent lab study showed that the new approach enables low anxiety recording and text-like editing of speech comments.

6.1 Preliminary deployment[9]

We first recruited a pilot class where we can specify the immediate requirement for repurposing RichReview as a peer discussion tool. This study allowed us to examine a different communication pattern encountered in education: students receiving and maintaining awareness about comments from several different peers (Gutwin et al., 1995). Also, in this follow-up study students produced multimodal annotations in addition to consuming them. Finally, the system we deployed in this study supported text annotations, which enabled us to observe differences in how students perceived the production costs between textual and non-textual annotation modalities (Marriott & Hiscock, 2002).

6.1.1 SYSTEM CHANGES TO SUPPORT PEER DISCUSSION

Moving from the instructor feedback setting to peer discussion demands a major change of system for enabling students to create multimodal comments in addition to being able to consume them. In response to user needs, we built a new version of RichReview by porting the legacy system to new hardware and software platforms. For clarity, we will refer to this new system as RichReview^{++}.

The interface of the previous system was designed for tablet hardware. However, it would be too costly to provide tablets to every student in the class. This led us to take the bring your own device (BYOD) approach that leverages the students' laptops and

[9] The text of this chapter was derived from a previous publication (Yoon et al., 2016).

desktops as a readily available deployment platform. This necessitated modification of the touch/stylus-centric interfaces, to one that is keyboard- or touchpad-centric. To create deictic gesture, we let the user drag the mouse pointer over the document, as the mouse was originally designed to point at something on the screen. For notetaking, we substituted free-form pen writing with typed commenting as shown in Figure 25, because writing ink strokes using the mouse input was inefficient.

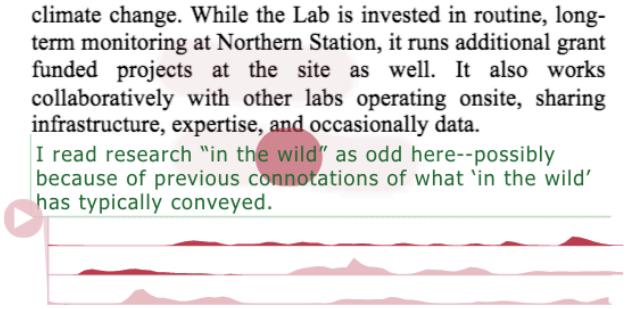


Figure 25. RichReview^{++} screen shot showing a thread of multimodal annotations containing text, voice, and gestures. In this figure, the red user created a voice + gesture comment in response to the green user's text comment.

The software-side update prepared the system for cross-platform and cross-browser access, because students were using a variety of machines. Our solution was to build RichReview^{++} as an HTML5-based web application that runs identically regardless of the user's platform or browser. To support voice recording on a web application, we exploited Media Capture API (getUserMedia) that offers access to the microphone input from the web app. However, this version of the system was missing the ability to edit voice comments because the speech transcription engine was too heavy to run in the browser.

Another important change the user needs is a transition from the single-user setting to the multi-user setting. For this, students in the new system could use the multimodal

threading feature where one can make an in-line comment in response to the existing comments as shown in Figure 25.

The presence of peers in the shared discussion space raises a privacy concern. Also, the literature suggested that support for private annotations would be important as a staging step before creating public comments and also to support active reading (Marshall & Brush, 2004). Consequently, our system added the ability to create private notes that only the creator could view and edit. To make the private notes visually distinguishable from the public comments, we extended them over the page boundary, creating a clear visual contrast (see Figure 26).

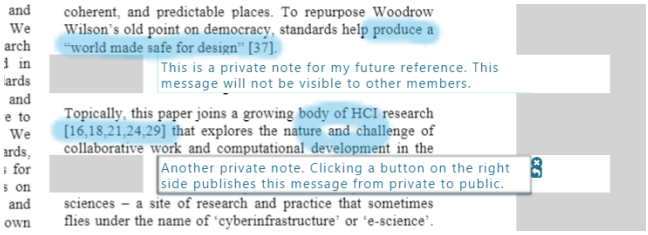


Figure 26. Private notes and highlights. Private notes extend beyond the page boundary for a clear visual distinction.

When many users create many comments, it becomes crucial to support efficient navigation of the existing comments. To enable quick and exhaustive browsing of the comments from multiple peer users, we added a comment history feature. This widget shows a chronologically sorted list of icons representing all of the comments present in a document, as shown in Figure 27.

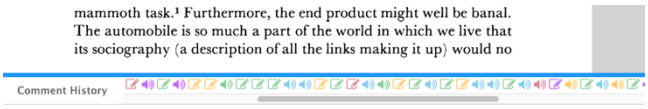


Figure 27. Comment history feature. A user can click one of the chronologically sorted links to existing comments to jump to the relevant page and the selected comment is highlighted.

6.1.2 DEPLOYMENT SETTING

The RichReview^{++} online discussion system was deployed to a graduate level social science course in the spring 2015 semester. In the class, an instructor taught 18 students, with 4 of them connecting from a satellite campus to the main class via a videoconferencing system. This seminar type class centered on individual readings and class-wide peer discussion activities, which were open-ended and student-directed. Assigned readings for a given week were first discussed online for a week, and then in an offline class-wide discussion session lasting 2.5 hours. Online discussion contributions made up 20% of the grade. The reading materials were composed of either 4-5 different conference papers, 3-4 chapters of a textbook, or a mix of both, totaling 150-250 pages per week.

Students used RichReview^{++} in two two-week long deployment blocks (weeks 6-7, and weeks 11-12 of the semester). At the end of each block of the study, we conducted a 30 min-long semi-structured interview with a focus on the way multimodal annotation supported discussion activities.

6.1.3 PARTICIPANTS

14 of the 18 students (4 females) participated in the study. The participants were mostly graduate students (1 undergraduate) in their mid-20s (M = 26.5, SD = 3.5). The students' familiarity with the course topic occupied the entire spectrum from novice to

very familiar. 5 participants were native English speakers, and the others spoke English as a second language. Students' proficiency in English ranged from intermediate to fluent. We placed the 14 participants into 3 groups with 4-5 members each in order to limit the volume of annotations on each group's document. The main reason for doing this was to make sure that RichReview^{++}'s visually rich comments did not take up all of the available screen real-estate. The student groups were balanced regarding background, gender, English proficiency, and campus location.

6.1.4 RESULTS

In the first block of the study, we observed 251 textual comments and 90 voice comments. Of the voice comments, 37% contained pointing gestures. Students produced 322 comment threads and 306 of these consisted of a single comment (Mean word counts = 31.8, Mean recording duration = 19.6). 16 comments had replies. However, the resulting threads were only 2-3 comments long and did not have much back and forth conversation. Three participants created the majority of their comments using the recording feature, three others made mixed use of modalities, and the rest created their comments using typewritten comments. In the second block of the study, participants made a similar number of textual comments (236) but much fewer voice comments (8) than in the first block.

Pointing Gestures

The pointing gesture seemed to be most useful as a visual aid for directing the listener's attention to locations on the page to which the voice recording referred. Out of the 90 voice comments created in the first block of the study, 37 employed the use

of pointing gestures, 25 of these referred to a single place of the body texts, and 12 pointed at multiple places (MAX=4, M = 1.41, SD = .77). One interesting property of annotations that referred to different locations was the shift in location granularity. For example, P7 moved between pointing to a phrase with a few words (e.g., "I don't want to agree with the statement that 'the most innovative ideas [gesturing over the phrase]' …") and pointing to the entire paragraph ("I agree with this view provided here in this part … [circling the paragraph]") in a single audio recording (Figure 28).

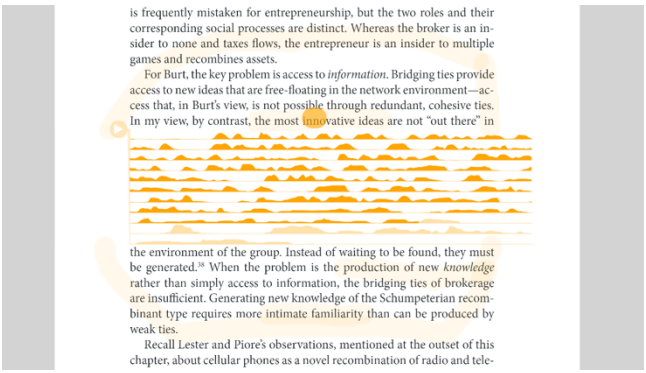


Figure 28. P7 referred to three different phrases and a paragraph using the pointing gesture feature.

Qualitative data also supported the benefits of pointing in tandem with voice. Students reported that pointing gestures were useful for referring to graphical element of texts (P1) and connecting multiple parts of the document in a single description (P7). To quote P1: "I found it particularly useful for pointing out things like diagrams… there was a few confusing diagrams and, that [the Spotlight feature] allowed me to draw out what was confusing." On the consumption side, students found gestures were helpful for understanding the speaker's intent (P9, 10, 11), which corroborates our findings in the first study.

Barriers to Creating Voice Comments

Students reported that creating voice comments (M = 3.08, SD = 1.38) required more effort than typing (M = 4.33, SD = .49). The reasons behind why our students felt it hard to record voice echoed previous findings of Marriott & Hiscock (2002). These reasons included lack of editing features, self-consciousness (non-native speakers concerned about their accents), and environmental constraints (e.g., working in a library or when a roommate was sleeping). Also, the linear and irreversible nature of voice recording made them feel compelled to keep speaking, which interfered with their thinking (5 of 11 students). The lower use of voice comments in the second block of the study can also be explained by the fact that students were busier at the end of the semester and had less time for creating multimodal comments.

6.2 Improving production of speech comments[10]

Two of our past studies as well as the previous studies in HCI literature suggested two major problems with speech comment production. First, speech commenters tend to be self-conscious during live-recording, because they were concerned that their comments might have speech disfluencies, such as 'um' or 'uh', stutters, or long pauses (Marriott & Hiscock, 2002; Scholl et al., 2006; Sivaraman et al., 2016; Yoon et al., 2016). People may also feel disturbed when hearing their own voice (Holzman & Rousey, 1966; Yoon et al., 2016). Second, editing of spoken speech comments is taxing. Over the past decade, several transcription-based speech editing systems have shown that using a transcript as a proxy for audio offers effective semantic editing of spoken content (Rubin et al., 2013; Sivaraman et al., 2016; Whittaker & Amento, 2004). However, these interfaces presumed an *a priori* audio transcription. To cope with the context of spontaneous speech commenting, these interfaces introduce separate modes of interaction: text-like audio copy and deletion, correcting live transcript from error-laden speech recognition, and re-recording for progressive revision (Sivaraman et al., 2016; Whittaker & Amento, 2004). Tracking and switching between these multiple modes can easily confuse users (Raskin, 2000).

This chapter suggests a solution to this problem of speech production. The key insight is that substituting the user's speech for a generic voice synthesized from a transcript can aid in reducing both (1) self-consciousness and (2) effort spent editing

[10] The text of this section was derived from a previous publication (Arawjo et al., 2017).

audio. In short, letting somebody else speak in-lieu of oneself reduces speech anxiety. Also, unlike previous interfaces, there is no need to re-record speech when making minor insertions or switch modes to perform different operations, because users can simply edit content with their keyboard. To explore the efficacy of these approaches, we built TypeTalker, a speech synthesis-based multi-modal commenting system.

6.2.1 TYPETALKER: A SURROGATED VOICE REDUCES SPEECH ANXIETY

We set the following three objectives to make the design decisions which embody the needs and challenges learned from prior work.

Reducing self-consciousness. A sense of being recorded can introduce anxieties in speech production. Two major sources of anxiety are the speaker's concerns about the way one's voice will sound to the recipients (e.g. 'um's), and the affective disturbance of hearing one's own voice. By using an automatically generated text-to-speech voice instead of the speaker's own voice, we let the user be less self-conscious about recording their voice.

Single-mode speech editing. Previous editing systems (Rubin et al., 2013; Whittaker et al., 1993; Yoon et al., 2014) maintain a loose correspondence between text token and source audio snippets. This setting (1) requires producers to do "double-work" for editing audio and correcting transcription errors, (2) introduces confusing interaction modes between audio editing and caption editing, and (3) slows down frequent and small edits (e.g., adding the past-tense ending, "ed"), since new words can only be inserted by speaking. In TypeTalker, synthesized audio is generated

based on text tokens as the user edits them. This tight coupling of edited audio to text tokens enables a unified single-mode revision process for audio (Figure 29).

Retaining expressivity of the original speech + gesture recording. A pure synthesized voice generated based only on the transcribed text misses the richness of multimodal inputs, such as pauses in the user's speech or co-expressive gestures. In TypeTalker, the synthesized speech retains expressivity of the source speech, such as natural pauses. Also, time-synchronized gestures recorded with the original speech are transferred to the corresponding words of the synthesized voice.

TypeTalker is a speech synthesis-based multimodal commenting interface. In TypeTalker, the user's voice entry is transcribed to later be synthesized into a computationally refined generic voice. As we show, the synthesized voice reduces speaker anxiety, since the audio in the standardized voice lacks the linguistic glitches of the original speech, and doesn't cause the affective disturbance of hearing one's own voice (Holzman & Rousey, 1966; Marriott & Hiscock, 2002; Yoon et al., 2016). In addition, we show that this approach reduces editing time by enabling an error-laden text transcription to act as a proxy for simultaneous editing of audio and any underlying metadata. This method allows temporal metadata, in this case gestures on a document, to be captured and replayed in sync with a speech comment, even after the comment is edited.

In TypeTalker, the user's voice entry is transcribed to later be synthesized into a computationally refined generic voice. As we show, the synthesized voice reduces speaker anxiety, since the audio in the standardized voice doesn't convey all the linguistic glitches of their original speech verbatim, and doesn't cause the affective

disturbance of hearing one's own voice (Holzman & Rousey, 1966; Marriott & Hiscock, 2002). A potential tradeoff of these psychological vantage points is that the synthesized voice loses multifaceted auditory sensations, including volume, inflection, and timing, all of which help to communicate nuanced and subtle semantics (Chalfonte et al., 1991). As a solution, TypeTalker *retains* some temporal richness of audio by transferring pause timings from the original speech to the synthesized speech. In addition, TypeTalker allows temporal metadata, in this case gestures on a document, to be captured and replayed in sync with a speech comment, *even after the comment is edited.* TypeTalker accomplishes this through an algorithm that respects the temporal alignment between the edited speech-as-text and the peripheral, extra-modal richness of the original recording, such as speech pauses and co-expressive gestures adopted from the RichReview system (Yoon et al., 2016, 2014).

The functional benefit of TypeTalker's synthesis-based approach is its streamlined workflow for revising speech (Figure 29) in comparison with that of traditional transcription-based speech editing systems (Figure 30). While the captions (text) in a traditional interface only work as placeholders for utterances, TypeTalker guarantees a match between audio and text because the voice is synthesized from the text. This approach enables simpler and more efficient revision, as caption correction and content editing are unified through single-mode keyboard editing over the transcribed text. For example, TypeTalker supports generative editing operations (e.g., insertion of new words or syllables) via simple text editing, while the traditional approach requires recording (or re-recording) additional voice content.

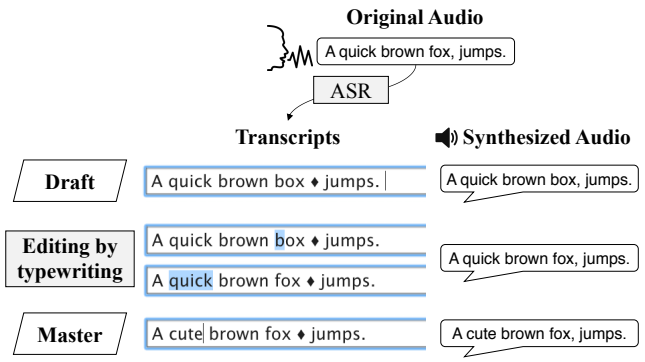


Figure 29. TypeTalker workflow. A user can finish different types of editing job in one-pass: correcting the ASR error ('fox' mistranscribed as 'box'), and changing a spoken word content (from 'quick' to 'cute').

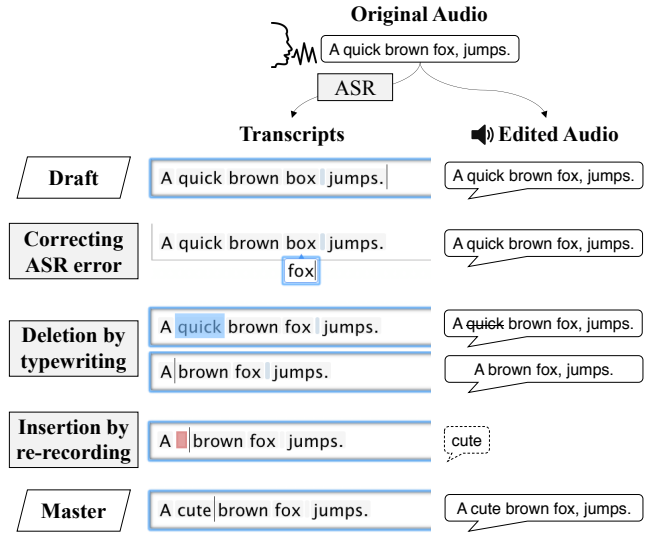


Figure 30. The traditional transcription-based speech editing workflow requires a user to switch between three different input modes: correcting captions ('box' to 'fox'), editing contents (deleting 'quick'), and re-recording (adding 'cute').

6.2.2 DESIGNING TYPETALKER

This section illustrates the user workflow using the TypeTalker interface. Imagine that a user wants to comment on a diagram. In our example, she begins a new comment beneath the prompt. A tooltip waveform and icon blink to remind the user that they are in recording mode (Figure 31. (a)). While recording, she can refer to

areas of the diagram by making deictic gestures (see (b)). Inside the text box, a red marker pulses at the place of insertion, reminding the user that they are in recording mode (c).

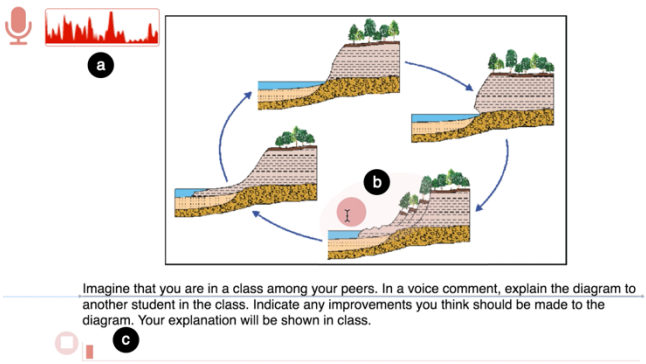


Figure 31. TypeTalker inside the RichReview system is designed to record, edit, and replay speech + gesture comments.

Once the user stops recording, the system is ready to support editing (Figure 32). It first swaps the blinking marker with the final ASR transcription results (a). In TypeTalker, each word of the ASR transcript is linked to time-stamped metadata (such as a gestures). To enable text-like single-mode editing, the system presents the transcription as a normal text inside the textbox by managing this audio correspondence data in the background, hidden to the user.

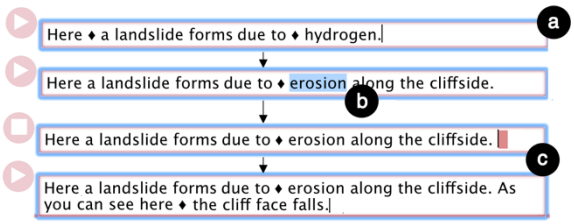


Figure 32. Editing process for the spoken comment

At this point, the user can review their comment and edit it through standard keyboard-based text editing. It is important to note that one can both correct

transcription errors and insert new content in the same textbox with a seamless and consistent keyboard interaction. For instance, in (b), the user fixed the mis-transcribed ‘hydrogen’ to ‘erosion’ as well as typed-in “along the cliffside”). Editing can also include deletion of portions of speech, punctuation revision, and pause manipulation. When the user wants to add new contents with gesture or speech pauses, they place their cursor at the end of the text and press the ‘Enter’ key to begin a new recording (c). The same revision process follows.

Upon pressing the ‘Play’ icon for playback, the system narrates the newly edited text, together with an animated visualization of gesture and pen input properly synchronized. This multimodal replay gives a vivid multimodal rendering that supersedes just reading the transcribed text.

One of our goals was to retain some of the expressive quality of the user's original voice. Of particular concern was the fact that speech-to-text lacks the user’s natural breaks in speech. We alleviated these concerns by transferring pause from the original speech to the synthetic voice. To help users control which pause they would like to keep, the in-line markers ‘♦’ denote a short pauses, while we append a period ‘.’ for longer pauses. We initially also transferred select pitch contours per-word according to their root-mean-squared-error with the synthesized word's contour; however, we were not able to find the right balance between the articulated prosody of speech-to-text voice and the user's natural prosody. The system used in our evaluation had the pause transfer feature only, but more than half of the participants reported that the retained pause timing could effectively convey the majority of expressive richness in the original speech even without having prosody transfer.

6.2.3 IMPLEMENTATION AND TECHNICAL DETAILS

We built TypeTalker as an interface extension of the RichReview^{++} system to exploit the benefits of the previous system, such as inline commenting and combined voice and gesture. Still, implementing the new synthesis-based design required additional works to apply the concept to the real world, as follows.

Real-time Transcription

For transcription, our system streams microphone data to the IBM Watson service (IBM, n.d.), which we also use for synthesis. For each recognized word we obtain timestamp data. As the system revises its matches, we store the current best match, and permanently append the final matches at the current insertion point when the user stops recording. Any special markers received from Watson are ignored.

Our early prototype opted to show the live transcription results in the textbox while recording; however, we observed in the first pilot study that on-the-fly transcription errors distracted users from their commenting job, because they became concerned with spotting and fixing errors as the transcript updated. Therefore, we opted to minimize distraction by presenting a simple blinking marker only, much like a text-editing caret.

Mapping Gestures to the Edited Tokens

In TypeTalker, gestures recorded during speech must be automatically remapped when the user edits the text. Since the synthesized voice is spoken at a different rate than the user's, gestures made during recording also need to be stretched. Both features rely upon situating words in the synthesized audio. Since timestamp

information was not available for the speech-to-text output, we ran the audio and edited transcript through HTK forced alignment (Cambridge, 2016) to obtain timestamps. From both the edited tokens and the speech-to-text timestamps, we then computed a sequence of "synthesized" tokens. Gestures could then be recomputed from both the edited tokens and the synthesized tokens through a split-map-reassemble process similar to Golovchinsky and Denoue's visual segmentation scheme (Golovchinsky & Denoue, 2002). Our method respects the time correspondence of speech tokens and gesture strokes. First, in the splitting phase, gesture strokes in the original recording (Figure 33. (a)) are chopped into pieces of strokes (see (b)), where temporal information corresponds with co-occurring tokens in the edited sequence. The chopped pieces are then mapped to the corresponding speech tokens in the synthesized sequence (c). Finally, we combine the reassembled gesture-piece sequence with potentially clipped pieces by lumping consecutive runs of gesture pieces into a single continuous gesture stroke (d) that respects the new beginning and ending time-stamps.

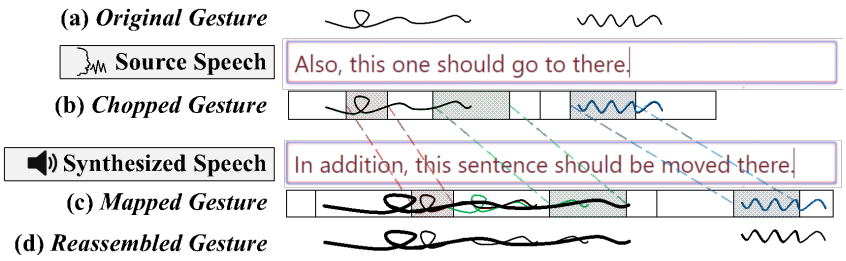


Figure 33. The split-map-reassemble process for gesture transfer.

6.2.4 STUDY DESIGN AND PROCEDURES

Our primary evaluation aimed to study whether our new design approach of TypeTalker could reduce producer speech anxiety and promote faster speech editing by comparing it to the SimpleSpeech system (Sivaraman et al., 2016), a design based on the previous approach. To draw out a quantitative comparison as well as qualitative implications for the future design improvements, we employed quantitative-major embedded design mixed methods where a task-driven lab study embeds exploratory qualitative inquiries such as observation notes and interviews (Creswell et al., 2007).

Participants

For this formative evaluation process, we recruited 15 young (18-22 years old, 14 female) undergraduate students at a US university. We selectively sampled participants who speak native or fluent English (12 native speakers), since the speech recognition system was optimized for standard American English pronunciations and accents. Our participants had different majors spanning across art, science, and the humanities.

Data collection and analysis

To set up a concrete and substantive use context, we put participants in the shoes of a student who takes part in the discussion activities of online coursework at a University. More specifically, we gave participants a series of commenting tasks that asked them to record their speech and gesture on given diagrams. The diagrams depict middle-school level academic topics, such as the ‘bottle recycle process’ or the ‘coastal erosion process’ shown in Figure 31. We only selected diagrams with very

easy concepts and minimal text, because we wanted the participants to focus on our interface rather than spending too much effort thinking about what to say. Each participant performed 3 sessions of tasks; in each, they created a paragraph-long speech comment. These tasks imposed proper amounts of effort on the participant to the extent that they had to leverage the full functionalities of the system in a reasonable time range (total 3~6 min) for this 1.5 hour-long study.

After the sessions for each condition, participants answered a set of surveys for rating perceived public/private speech anxiety and overall task loads. The anxiety measure was adopted from the Scheier & Carver's Self-Consciousness Scale (SCS-R) by selectively contextualizing four questions about public speech to the asynchronous speech recording use case (Scheier & Carver, 1985). The workloads were measured as the weighted NASA-TLX scale (Hart & Staveland, 1988). Participant activities were also logged to measure the number of different recording and editing behaviors as well as the transcription results.

To collect the qualitative data, the investigator sat behind the participants' workbench, observed her task practices, and took notes of any notable incidents. Implications from the observations were referred back from the post-task interview for two purposes. First, we asked 'how' and 'why' questions to the participants to better understand the rationale behind their behaviors (Lofland & Lofland, 1971). Second, we did member checking (Maxwell, 2013) to validate our on-the-fly interpretation of participant behavior.

We performed the paired t-test for the quantitative data, such as the self-consciousness or workload indices. For generating quantitative implications, we

conducted theoretical sampling (Glaser & Strauss, 1967) by comparatively analyzing data from the two different UIs. After collecting and transcribing interview data into texts, the lead investigator performed an open-coding followed by a flat-coding to draw out theoretical categories of the implications in consultation with the coauthors. To maximize the validity of our findings, we triangulated different types of data, and consistently looked for negative cases to falsify potentially defective evidence (Maxwell, 2013).

6.2.5 RESULTS

The participants generated a total of 90 comments (15 × 2 × 3) for the tasks. On average, the comments were 20.3 sec long (SD = 16.1) with 39.6 words (SD = 36.2) for TypeTalker, and 18.0 sec long (SD = 14.3) with 36.9 words (SD = 35.4) for SimpleSpeech. They also made a couple of gesture strokes for each session (M = 1.50, SD = 2.08). None of these basic measures were significantly different between the two conditions.

Average recognition accuracy of the source speech measured as word error rate (WER) was .19 (SD = .10) in the TypeTalker condition, and .16 (SD = .09) for the SimpleSpeech condition which are slightly higher than IBM's official data of .104 (Soltau et al., 2014) possibly due to the participants' speech disfluencies. The response speed of the transcription engine was near real-time as we live-streamed the audio to the Watson's cloud server in the 16-bit 22.05 kHz PCM format.

Reduced Self-consciousness

The participants perceived significantly less public/social self-consciousness during speech when using TypeTalker, thanks to the synthesized generic voice imposed less concerns about public performance than the peer system that records audio as is (see Figure 34, left). In our SCS-R measure, ratings for public/social-anxiety were significantly lower with TypeTalker (M = 2.57, SD = 1.13) than SimpleSpeech (M = 3.06, SD = 1.05, p = .019, paired t-test, Cohen's d = .46).

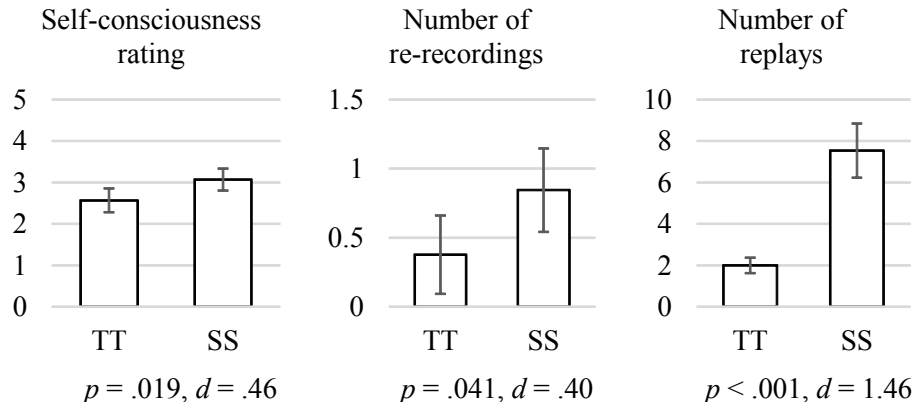


Figure 34. Quantitative results from TypeTalker (TT) and SimpleSpeech (SS) conditions (95% confidence intervals).

From the qualitative responses, we found that a total 12 of 15 participants reported lowered self-consciousness. First, 7 participants (P1, 2, 4, 8, 10, 13, and 14) reported that the TypeTalker interface alleviated their concerns about the way their speech will sound to the recipient, because the machine's voice doesn't retain small speech disfluences including 'uh', 'um', stutters, hesitations, or long pauses. In contrast, voice recordings in SimpleSpeech made them "nervous that I would just keep going like 'um, um...' in the middle of my statements. It just seems like there was a lot more that

could go wrong that way. (P13)". There was no participant in our sample who felt more anxiety in the TypeTalker condition.

Second, 7 participants (P2, 4, 5, 6, 7, 11, and 13) liked that they don't have to listen to their own voice, which often causes affective disturbance (Holzman & Rousey, 1966). To quote P13, "I personally hate listening to my own voice on recordings [laugh]. It's weird to me. It's a little off-putting (P13)." This disturbance remained salient during the revision phase, as P11 stated "I don't like hearing my own voice. So when I try to replay them, I almost muted the computer". The other 8 participants didn't mention the affective disturbance from hearing their own voice.

Finally, 6 participants (P4, 5, 8, 9, 13, and 14) were less concerned about making mistakes while using TypeTalker because they "knew that it was easier to correct those mistakes (F5)" and "required much less work (F8)" using the keyboard interface afterwards.

Effective Revision

The participants were unanimous that TypeTalker's type-written editing was not only easier to learn, but also more lightweight and effective for editing. The participants liked familiarity of the normal text editing interface, single-mode, and no need for re-recording. This quote summarizes the implications well:

> *"TypeTalker was easier, just because it was very similar to like normal typing, I could just go in and fix things, and you know I could change words, I could change sentences if I wanted to without having to worry about it, whereas with SS, if I wanted it to change or rephrase something, I have to go*

in and re-speak it, and it usually comes out sounding different than like louder, like just awkward (P14).”

Reduced confusion from unified editing mode. A total 9 of 15 participants said that TypeTalker’s single-mode editing was straightforward to learn and use, and more efficient. Although most of them felt that SimpleSpeech’s interface was easy enough to get accustomed to in a reasonable learning time, they oftentimes felt the interface “confusing (P6),” “frustrating (P8),” or “not knowing what to press (P5)”. They felt that there were too many options, and as P12 stated, “when to say, when to type, when to press enter was different than expected”. This suggests that TypeTalker’s design decision to unify audio and text editing significantly improves editing efficacy. For some users, audio-text division of SimpleSpeech was not only about the mode confusion issue during editing. Sometimes, when users accidentally deleted an audio token, they tried to recover it by typing that transcription to the nearest token, losing audio although showing the correct transcript: “I often deleted my voice recording closing a glitch. That was the hassle (P10).”

Efficient content editing. Editing spoken content was more efficient in TypeTalker, because, unlike in SimpleSpeech, it didn’t require participants to re-record parts (11 of the 15 participants). When asked about how hard they worked for editing content other than transcription errors, the survey response showed a trend that they edited more content in TypeTalker than SimpleSpeech ($p = .082$, Cohen’s $d = .47$). More pauses were edited during the revision process as well. In the original speech data, the number of pauses were not significantly different, but the end results had significantly less pauses in TypeTalker condition. This is possible because of a

trend where more pauses were deleted during editing in the TypeTalker system ($p = .093$, Cohen's $d = .33$).

To insert new audio contents in SimpleSpeech, users had to re-record the part of speech, since there is no way to create new audio from the edited text. For our participants, this re-recording worked as the major drawback for editing content (11 of 15 participants). This implication is reflected in log data that shows that users rarely re-recorded in the middle of a SimpleSpeech stream ($M = .84$, $SD = 1.19$). Also, when they were given the typing capability to add contents in TypeTalker, the use of the insertion feature became significantly rarer ($M = .37$, $SD = 1.12$, $p = .041$, Cohen's $d = .40$), because they preferred to just type to insert voice rather than recording that part again. They preferred not to re-record speech not only because it was cumbersome, but also because the inserted voice sounds awkward and felt "forced in (P5)", "misplaced (P8)", "louder (P14)", "off-flow (P14)", or "choppy (P15)".

Implications on correcting transcription errors. Even though there seems to be more pressure to correct transcription errors in TypeTalker than SimpleSpeech "because the program will specifically read what was transcribed (P8)", such pressure was evened out by the three factors beneficial to the TypeTalker condition. First, editing by keyboard input was easier in TypeTalker as stated above. Second, SimpleSpeech users also felt pressure to fix mis-transcriptions, so that the recipient of the messages wouldn't be confused by the wrong text (13 of the 15 participants). Finally, editing transcriptions in SimpleSpeech forced the participants to re-listen to their audio, because they had to match the text to the voice. There were significantly more replays in SimpleSpeech than TypeTalker ($p < .001$, Cohen's $d = 1.46$). Re-

listening was upsetting for the participant, not only because it was cumbersome (P2, 7, 8, and 14), but because re-listening during editing (P11) also caused the self-disturbance problem of hearing one's own voice.

Nonetheless, some other users liked that they could re-listen to their voice in SimpleSpeech, because it helped remind them of the content of their original narration (F9 and 11). Although the TypeTalker system didn't have the re-listening feature for replaying the original voice, one might improve the transcription correction process by including an in-situ replay feature as a mnemonic device for the system in the future (e.g., replay the snippet of original speech, when selecting words for editing).

Valued Richness of Original Audio

8 of the 15 participants liked TypeTalker's pause mark feature that transfers subtle timings from the original voice to the machine's voice. The pauses in the machine generated voice could make it "sound more human-like (F5)", and enabled them to verbally "emphasize (F2)" a phrase by generating some temporal suspense. Also, listing items such as "Croatia <pause>, Slovenia <pause>, and all (F2)" sounds more natural with having the pauses in-between. This implies that future machine-synthesized voice technologies can largely benefit from transferring richness of the original voice to the synthesized voice.

Although all producers admitted that the machine's voice reduces speech anxiety and enables efficient editing, 8 missed rich acoustic expressions from their own voice, such as "emotion (P8)" or nuances (e.g., "sarcasm (P9)"), delivered by subtle "inflections (P4, 6, 14)". A few (P3, 4, 14) also wanted to retain the identity of the

original speaker (e.g., "gender (P4)"). Producers may have been concerned that this loss of expression would impact the recipients of their comments. To explore how comment consumers were affected by the machine voice – whether they, too, missed natural expression, and to what extent – we conducted a follow-up qualitative study, described in the next section.

6.2.6 CONSUMER-SIDE EVALUATION

The goal of this follow-up study was to understand how the content-consumer's comprehension and experience are influenced by the two types of voice comments generated from each interface: TypeTalker with a machine's voice, and SimpleSpeech with a human voice. For this study, we collected qualitative data from participants who conducted a set of consumption tasks on the comments produced during the first study.

Sampling

We recruited 10 (19-39 years old, 6 female) participants at Cornell University. We diversified the consumer demographics by recruiting participants from varied academic backgrounds. Also, unlike the primary evaluation, 4 were non-native English speakers comfortable in written English. None of them had participated in the producer-side evaluation.

Tasks

To mirror the task context of the primary study, we placed participants in the shoes of a student in an online peer discussion context, and asked them to critique producer-generated explanations of various diagrams, focusing on audio delivery. Specifically,

we let them first *listen* to each speech comment, and then *type* a short response (2-4 sentences) evaluating each of them. We explicitly asked them to play the audio rather than reading the transcribed texts, so that they could listen to the comments in order to compare generic and human voices.

Procedure and Materials

At the beginning of the study, the investigator gave a brief tutorial about how to use the interface to create a text response, then began the first session. There were a total of 2 sessions, one for each interface condition (TypeTalker and SimpleSpeech) which lasted a total of ~45 minutes. In each session, the participants conducted 2 commenting tasks that took ~5 min each. Each task contained a comment randomly assigned from a producer in that condition from the primary study, with the constraint that no diagram was presented to each participant more than once. Condition order was counter balanced. After both sessions, the study was concluded with an audio recorded, ~10 min-long semi-structured interview.

Results

Consumers were ambivalent as to preference of voice type. 5 (C2, 5, 6, 7, 9) preferred a human voice in general, but were also not particularly bothered by the machine voice. 4 (C1, 4, 8, 10) did not express a clear preference for either voice, and one, a non-native speaker (C3), preferred the machine voice. Even those consumers who preferred the human voice did not find that the machine voice hampered their comprehension. For instance, C2, who preferred the human voice, stated, “The voice, although robotic, was concise and it was relatively easy to follow along with the

diagram." C5, who also preferred the human voice, said that "the machine voice itself didn't bother me. It was fine." 2 participants (C1, C10) did not even notice there was a difference between conditions until pressed.

In general, consumers cited improved elocution through standardization as a major benefit of the computer generated voice, with possible trade-offs of lost expressivity and engagement. For example, C2 thought the machine voice was "easier to follow because it's easier to understand a slow robotic voice," while C8 found the machine voice preferable "if someone has an accent, or speaks really fast or slow," and remarked that a standardized voice "would be helpful for a wider range of students." Awkward lengthy pauses, disfluencies, and speaking rates were continually cited as an issue for human-voice comments. C3 explained, "it's much more comfortable to listen to the machine voice for me. Because the human voice, they have pauses and they [speak] more slowly." Even C9, who preferred a human voice in general, admitted, "the [human voices] had a lot of awkward pauses. That does follow the natural way of speaking, but because of that, it also was more difficult and unclear. There's a lot of rapid changes in pace. So it's like a pro and con."

Nevertheless, some consumers felt that the benefit of a standardized voice also imposed an effect on their engagement. C9 went on to state, "when I am listening to a machine, it is a little harder to engage […] Because it's like a one-tone voice, and one-speed." In addition, C7 found a human voice more "soothing" and "easier to pay attention to" than the "monotone" machine voice. Future work to transfer prosodic features could remediate or remove this drawback.

TypeTalker's improved text output was also appreciated. 9 out of 10 participants found the text useful to their comprehension of the comment, especially for reminding them of the audio (the last did not mention the text). In the human condition, 4 participants cited issues with SimpleSpeech-produced comments: two (C6, 10) noticed typos and were "a bit confused (C6)" as to the mismatch between speech and text, while the other two (C4, 8) complained about punctuation and grammar. However, two participants (C5, C1) in the TT condition also mentioned correcting minor mistranscriptions in their remarks to the commenter, which highlights the addition burden placed on commenters by ASR accuracy. Many users could not comment directly on the quality of the text as they found the quality of speech enough for their understanding.

6.3 Discussion

From the findings of our evaluations, we gleamed several additional needs and considerations for the design of future speech editing interfaces.

Implications of a trade-off between the human voice and the standardized voice

Most of participants in the primary study felt that the synthesized voice sounds more professional whereas the human voice has a better personal touch. They thought that the professionalism of TypeTalker comments would be better received in formal or official settings (e.g., lecture, audio book publication, etc.), while for other settings, such as snapchatting or a lecture in a smaller class, use of their own voice would fit better, since it can convey character and personality of the speaker. Interestingly, consumers in our follow-up study interviews made similar remarks. This implies that

the use-case of the speech commenting system might be one of the major factors in deciding on which way to present spoken comments.

Results from the consumer-side study suggest that standardized voice might enhance the listener's comprehension by reducing distracting aspects of speech such as disfluencies, lengthy natural pauses, and fast speaking rates. This effect was particularly noticeable for non-native English speakers, because for them, comprehension of speech was a priority. This implies that TypeTalker may enable a more diverse group of individuals to hold a discussion than that accomplished by recorded voices alone. This would be especially useful for a multi-cultural CSCW context. However, more work needs to be done to improve the naturalness and expressivity of the machine voice, in order to tackle the trade-off from the loss of personal touch.

Retaining more personal touch from the original speech

Even though the transferred pause timings could convey temporal subtleties of speech, such as rhythm and suspension for emphasis, other acoustic qualities remained untouched. For example, transferring natural intonation, prosody, and loudness of the original speech could make the synthesized voice sound more similar to the original. The speech synthesis research community has been presenting a set of bedrock technologies, such has pitch-synchronization (Valbret et al., 1992) or emotional prosody modelling (Schröder, 2001), that can be used to realize such features.

Hybrid approach: mixing original and synthesized voice

Future designs could take a hybrid approach that takes different advantages from both of the approaches by acoustically mixing the synthesized voice with the original speech. This could enable type-written generation of audio without having to re-record that part of speech. The latest speech conversion technique promises seamless stitching of synthesized voice into an existing speech stream (Jin, Z., Finkelstein, A., DiVerdi, S., Lu, J., and Mysore, 2016). Also, a speaker de-identification technique can be used when users want to anonymize their voice for reduced self-consciousness (Qin Jin et al., 2009).

6.4 Summary

From the field evaluation of RichReview against peer-discussion activity in the classroom, we learned that the major barriers to using the multimodal discussion tool were threefold: editing speech was effortful, live recording caused speech anxiety, and students do not want to hear their own voice. To solve this problem, we took a new approach that substitutes the user's voice with a mirror narration from a speech synthesizer. Based on this transcription + re-synthesis approach, we designed and implemented a new interface called TypeTalker as a part of the RichReview system. We conducted a comparative lab study for TypeTalker to test its efficacy in speech comment production in comparison to the conventional transcription-based speech editing interface. The results confirmed that TypeTalker reduces speech anxiety, mitigates self-affective disturbance, and accommodates text-like editing of speech, especially for the insertion of new content.

This chapter's evaluation tested the efficacy of the TypeTalker solution for facilitating speech production workflow in lab tasks. The natural next step is to study real-world efficacy of our new approach. In the next chapter, to observe and evaluate how real users use TypeTalker for real peer-discussion tasks, we present the design and results of a TypeTalker deployment study in a massive open online class.

7 MOOC deployment of TypeTalker for online peer-discussion

MOOCs have the potential to democratize quality education (Barber, Donnelly, Rizvi, & Summers, 2013; Selingo, 2013). The digitized educational content can be delivered for little cost to anyone with access to the network. However, the other side of distance education is that online students are physically alone, missing the opportunity for social and personal interaction that classrooms can provide. Many researchers have studied ways to support such personal learning experiences in MOOCs through peer discussion (Coetzee et al., 2015; Coetzee et al., 2014; Huang et al., 2014; Mak et al., 2010). This chapter extends these previous works and our new multimodal peer discussion system, introduced in chapter 6, to provide face-to-face like rich peer learning experiences to students in MOOCs.

In the previous chapter, we identified the problem that students feel self-conscious and find it difficult to create speech comments in rich online peer discussion, and presented an interface solution called TypeTalker. To test the real-world efficacy of the solution, we updated the interface of RichReview with TypeTalker, and deployed the new peer discussion system to a MOOC run from CornellX/edX.

7.1 Design and implementation

We built a new version of peer discussion system (RichReview.net) for MOOCs. In a MOOC, anyone can enroll in the class, and then use the peer discussion system

anywhere, anytime. In response to such open and diverse user environments, our updates were focused on supporting the rich discussion features at scale in the field.

Implementing the TypeTalker interface as a part of RichReview was one of the necessities for supporting voice + gesture interaction, with reduced self-consciousness and anxiety, as explored in section 6.2. While scaling up, we noticed that the implementation of the previous system might yield very long response times for the pause timing transfer phase of the speech synthesis process. This is because running the cloud-based time alignment module could significantly delay the process due to the network bottleneck when there are too many requests. As a solution, instead of transferring the pause timings of original speech, this version added a short pause (0.4 sec) after each sentence of the synthesized speech. Other than this change in the pause time transfer, the core work process of speech transcription and synthesized narration remained the same.

We also built a group assignment module that automatically generates small-multiple discussion groups. The module allocated groups of 10 students by arrival order. This assignment method is known to yield heterogeneous groups thanks to the innate diversity of MOOC student population (Kulkarni et al., 2015). From the previous on-campus deployment study, we learned that groups of about four to five students did not have the critical mass for successful discussion because students reported that there were an insufficient number of peer comments that they can read or reply to. In this study, to hit the critical mass, we doubled the size of group to 10.

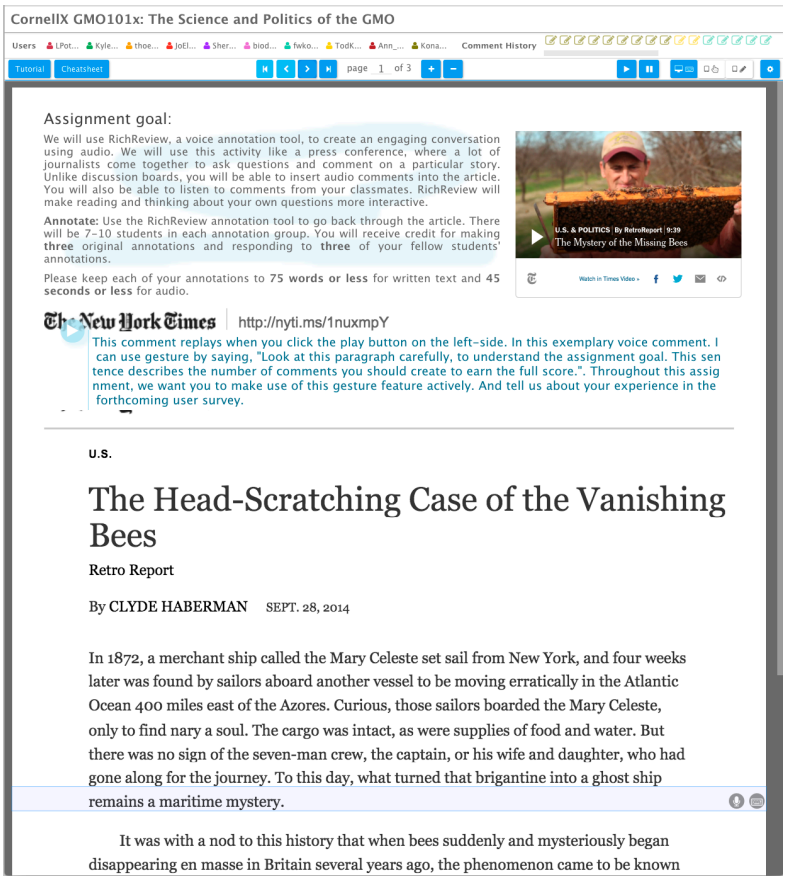


Figure 35. A screenshot of the RichReview discussion system running on a browser.

Finally, we revised a student data management module of RichReview to interface student data (e.g., profile, peer-group assignment, and credits) to the edX.org ecosystem. edX-compatible RichReview.net was built as a stand-alone node.js web application that talks to edX.org through Learning Tools Interoperability (LTI), a standard protocol for educational content and resources. The protocol provided secure access to the student data (e.g., fetching student profiles and their progress) and a programmable grading scheme (i.e., giving credits for the RichReview assignment). We preferred to run it through a standalone LTI interface rather than through edX's XBlock integration since it offered flexible and dynamic group assignments, which edX does not.

In addition, complying with the institutional policies of edX and FERPA regulations, the system could offer an accessible interface to the students with visual or auditory impairments. For example, at the recruiting phase of the study, the system automatically inquires whether the user is using screen reader software. Users who want to use accessibility features are redirected to the version of the system that conforms to Accessible Rich Internet Applications (ARIA) standard.

7.2 Study design and procedures

The goal of this study is to compare the impact of the rich online discussion for student learning with the conventional text-based discussion. To examine this core inquiry from various angles, we gathered student data on discussion activity, learning gains, and perceived efficacy.

Deployment setting

For the rich peer discussion activity, motivating students to participate was important because the students should not only listen to, but also generate multimodal comments. Hence, when choosing the target MOOC for the deployment, we prioritized the courses with debate-ready subject matter. With help from the Academic Technologies group at Cornell, we recruited a new CornellX course called GMO101x: The Science and Politics of the GMO, hosted by edX (CornellX, 2016). The course was designed to develop critical thinking and scientific literacy for understanding genetically modified products. In this course, a peer-discussion assignment could expose students to different perspectives and ideas. Considering the diverse educational backgrounds and academic interests of MOOC students, we selected a

short and easy-to-read news article as the discussion material: a 975-word long *New York Times* article, 'The Head-Scratching Case of the Vanishing Bees' (Haberman, 2014).

Experiment design

This study employed a between-subject design with two experimental conditions: voice-and-text and text-only. Students in the voice-and-text condition used the reference RichReview.net system described above. The system supported voice commenting with the TypeTalker interface. It also supported inline text annotation for typewritten comments, because it was imperative to offer a choice of modality for participants who do not want to use voice at all. For the baseline text-only condition, we built a voiceless version of the system by taking out the TextTearing feature. We chose to use the textual version of RichReview.net instead of using another conventional textual system (e.g., Piazza or nb) for the sake of a fair comparison where all subjects can use the same basic beneficial features, such as inline threading and fluid layout, regardless of their experimental condition.

Procedures

For this five-week course, the discussion assignment was given at the third week, which is when students could leverage the concepts from the previous weeks for the task while they still had a high level of motivation and engagement for the course. The students had the whole week to post original comments and replies. The first page of the discussion material showed the task description as in Figure 36. To motivate participation to the assignment, we gave 5% of course credit to students who started

three original discussion threads and another 5% to those who replied to three different students. We set a length limit for comments to motivate students to focus on quality over quantity.

> **Assignment goal:** We will use RichReview, a voice annotation tool, to create an engaging conversation using audio. We will use this activity like a press conference, where a lot of journalists come together to ask questions and comment on a particular story. Unlike discussion boards, you will be able to insert audio comments into the article. You will also be able to listen to comments from your classmates. RichReview will make reading and thinking about your own questions more interactive.
>
> **Annotate:** Use the RichReview annotation tool to go back through the article. There will be 7-10 students in each annotation group. You will receive credit for making **three** original annotations and responding to **three** of your fellow students' annotations.
>
> Please keep each of your annotations to **75 words or less** for written text and **45 seconds or less** for audio.

Figure 36. The task description given to the MOOC students.

Survey

To understand students' perceived efficacy, we conducted a survey right after the deployment week ended. The questionnaires focused on various aspects of student experiences of using RichReview for the peer discussion assignment (e.g., effectiveness for learning, sense of community, perceived usefulness). The surveys for the two experimental groups had the same set of generic questions, but the questionnaires for the voice-and-text students included additional items about the features of TypeTalker and their perception (e.g., speech recognition quality, speech

anxiety, etc.). For ratings, we used a five-point Likert scale that ranged from "Strongly Agree (5)" to "Strongly Disagree (1)." The respondents also filled out a typewritten questionnaire on their reasons for the ratings. These qualitative responses were analyzed using thematic coding based on grounded theory methods (Glaser & Strauss, 1967).

Recruitment

We recruited the students through an opt-in volunteer procedure. When they clicked the "Start the discussion" button of the third week's assignment page, students were prompted with an introduction page that describes the objectives and conditions of the study. The students who agreed to participate in the study were redirected to the RichReview version of the discussion system. In an educational tool like this, it is important to give everyone the same opportunity to receive the promised amount of credit, whether one participated in the study or not. The students who were either ineligible for the study (e.g., having hearing or visual impairments) or who disagreed with the terms and conditions of the study were redirected to an accessible version of the peer discussion system that has most of the basic features of the conventional threaded discussion tools to allow them to finish the task to get the course credit.

7.3 Results

A total 6424 students visited the online course. Of those, 173 (2.6%) finished the course. The low retention rate is a typical feature of MOOCs where there is significant attrition in student numbers through multiple stages of dropouts called the funnel of participation (Clow, 2013). Among those who finished, 147 students participated in

the study (67 females, 67 males, 13 unidentified gender). The participants were generally well educated (20 doctorates, 52 master's, 44 bachelor's, 17 high-school graduates, and 14 others). The two experimental groups were nicely balanced in terms of gender (46.8% females in the text-only group and 53.3% in the voice-and-text group) and education level (Mann-Whitney $U = 2230.0$, $p = .3758$).

For this one-week deployment, the 147 participants were assigned to 15 groups of between nine and ten students. Of those, 78 were assigned to seven voice-and-text groups, while the other 69 were assigned to eight text-only groups. For the analysis, we used only the discussion data from active users (i.e., students who created at least one comment). In the voice-and-text group, 55 active students made 236 text comments ($M_{length} = 23.1$ words, SD = 23.5) and 127 voice comments ($M_{length} = 25.3$ words, SD = 18.9). In the text-only group, 55 active students created 357 comments with an average length of 25.0 words (SD = 17.6). On average, a student generated slightly more than three original comments ($M_{\#\ original} = 3.25$, SD = 1.48) and three replies ($M_{\#\ replies} = 3.35$, SD = 1.97), because those were minimum criteria for getting full credits from the assignment. There was no significant difference between the experimental conditions regarding the word counts ($p = .27$) or the number of comments ($p = .42$).

Survey results. The respondents to the questionnaire included 22 of 69 text-only students and 31 of 78 voice-and-text students. Their ages ranged from 22 to 73 years old ($M_{age} = 37.6$, SD = 12.7). They were well educated (94.3% had a bachelor's degree at least) and good at English (the mean perceived English level was between

fluent (4) and native (5), $M_{English} = 4.26$, SD = .96), which implies that the demographic of the survey respondents was representative of the study participants.

As summarized in Figure 37(a), the group in the text-only condition rated different aspects of their discussion experiences higher than the voice-and-text students. The survey ratings from the two experimental groups were analyzed using a mixed-design ANOVA with a between-subject factor of experimental condition (2 levels, text-only and voice-and-text) and a within-subject factor of questionnaire (4 levels from "Helpful for learning," "Willing to use the system in the future," "Feeling that students care about each other," to "Feeling connected to peers," see x-axis). Mauchly's test indicated that the assumption of sphericity had been violated ($p < .001$), therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .62$). There was significant main effect of the experimental condition ($F(1, 50) = 8.08$, $p = .0065$, $\eta^2 = 0.08$) indicating that the ratings were higher in the text-only condition in general. The main effect of the questionnaire ($p = .13$) was not significant and there was no interaction ($p = .63$).

Text-major vs. Voice-major students. Although students under the voice-and-text condition could use both types of comment, each user had a personal preference to a type of mode. As shown in Figure 37, we categorized the students into two sub-groups of the voice-and-text group depending on which modality they used more: text-major or voice-major (i.e., a voice-major student is one who created more comments using TypeTalker than vanilla text). Fourteen of the 55 voice-and-text students were voice-major, which means that they created more comments using voice than text. Thirty-nine were text-major who created more text comments than voice comments,

and two others were balanced. Among the text-major students, seven tried using voice comments at least once, and the rest (32) never tried voice.

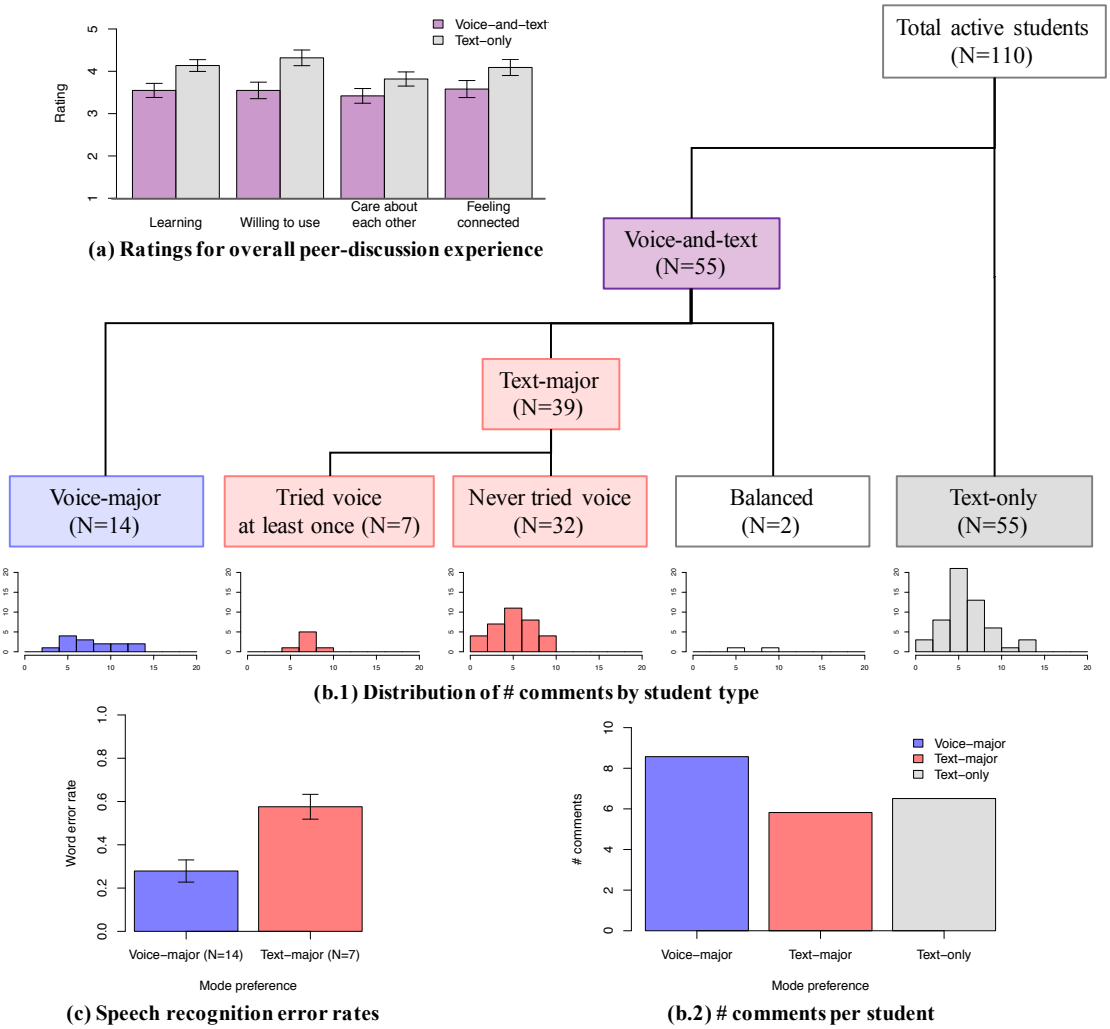


Figure 37. The students were classified by the experimental condition and then by modality preference. We compared discussion activities of different types of students as follows: (a) Students in voice-and-text condition gave a higher rating for the system than text-only students; (b) Voice-major students generated more comments than text-major students or students in the text-only condition, suggesting a higher level of engagement; (c) Speech transcription results were more accurate for the voice comments from the voice-major students than that from the text-major students.

We examined the number of comments created by the students in different groups as an indicator of students' level of engagement with the discussion activity. As shown in Figure 37(b), the voice-major group created significantly more comments ($M_{\#\ comments} = 8.57$, SD = 3.27) than the text-major group overall ($M_{\#\ comments} = 5.82$, SD = 2.30, unpaired t-test, $t = -3.41$, $p < .001$, Cohen's $d = 1.06$). This tendency is attributed to four voice-major super-posters who created more comments (>10) than any other students in the text-major group (the far right-end of Figure 37(b), blue). But the two groups did not show a significant difference in terms of education level ($p = .27$) or final grade they received from the GMO101x course ($p = .21$).

Speech recognition accuracy. To check if accuracy of speech recognition affects students' modality choice between voice and text, we compared word error rates[11] (WERs) of transcriptions from the speech comments of voice-major students and text-major students. As shown in Figure 37(c), there was a significant difference in the WERs for voice-major students ($M^{WER}_{text\text{-}major} = .28$, SD = 0.26) and the text-only students who tried voice ($M^{WER}_{voice\text{-}major} = .58$, SD = .29, $t = 3.33$, $p < .001$, $d = 1.09$). This suggests that some students first tried using voice, but converted to text as they judged that the quality of speech transcription was very low, to the extent that benefits of using speech fell behind the added efforts of correcting the recognition errors.

---

[11] Measured with the Levenshtein distance between machine transcription results and ground truth. Since it was too costly to generate ground truth transcriptions for every voice comment, we used the samples of the two latest comments for each student.

To better understand why voice-major students' narration yielded better transcription results, we examined their speech comments with a focus on the factors that can affect the accuracy. By listening to the actual recordings of students in the two groups, we could characterize recording quality and speech clarity as the major determinants of the transcription accuracy. Among seven students in the text-only group who tried voice, only two were native English speakers and presented quality voice recordings. The other five had issues of either thick, non-native accents which resulted in mis-transcriptions, or bad voice recording quality, such as wrong microphone position, background noise, or low volume. In contrast, recordings of voice-major students presented relatively better audio quality than that of text-only students. Also, 10 of 14 voice-major students were native English speakers, and the other four spoke English fluently.

Ratings for text vs. voice. From the voice-and-text responses, we could conduct within-group comparisons between the ratings for the two different commenting modes: text vs. speech (TypeTalker). The survey ratings were analyzed using a mixed-design ANOVA with a between-subject factor of experimental condition (2 levels, Text and TypeTalker) and a within-subject factor of questionnaire (5 levels from "Helpful for articulating my idea," "Helpful for understanding peers' idea," "Made me feel nervous," "Easy to understand by reading its text/captions," to "Easy to edit and revise," see x-axis of Figure 38). Mauchly's test indicated that the assumption of sphericity had been violated ($p < .001$), therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .50$). There was a significant effect of the experimental condition ($F(1, 50) = 8.08$, $p = .0090$, $\eta^2 = .063$). There was

an expected main effect of the questionnaire ($F(2.96, 171.51) = 4.95$, $p = .0027$, $\eta^2 = .04$). We also observed a significant interaction ($F(2.96, 171.51) = 4.22$, $p = .0068$, $\eta^2 = .03$). Hence, we conducted further within-group comparisons of rating for each individual questionnaire item.

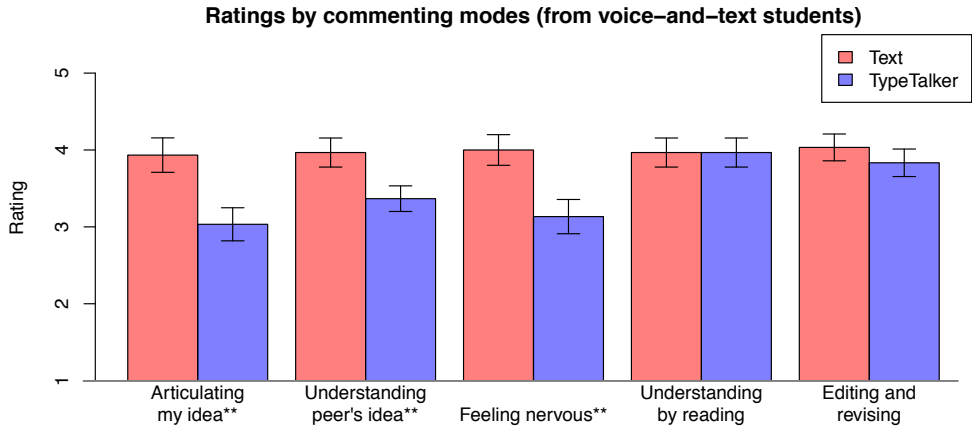


Figure 38. These ratings from the voice-and-text students depict perceived efficacy for the two commenting modes comparing text vs. voice commenting methods. (indicates the significance of $p < .01$. The error bars represent 95% confidence intervals). The higher rating is better (e.g., lower nervousness).**

For between-group comparisons of the individual ratings, we used Wilcoxon Z test with Bonferroni adjusted *p*-values. The respondents felt that they could better articulate their idea ($M_{text} = 3.93$, SD = 1.14 *vs.* $M_{voice} = 3.03$, SD = 1.10, $d = .80$, $Z = 78.5$, $p = .0014$), better understand peers' ideas ($M_{text} = 3.97$, SD = .96 *vs.* $M_{voice} = 3.37$, SD = .85, $d = .66$, $Z = 76.5$, $p = .001$), and felt less nervous ($M_{text} = 2.00$, SD = 1.02 *vs.* $M_{voice} = 2.87$, SD = 1.14, $d = .80$, $Z = 92.5$, $p = .0035$) when using text rather than voice. However, they felt that understanding peers' TypeTalker comments by reading its transcribed caption without replaying audio is as easy as understanding text

comments (p = .80). In addition, ease of editing and revision using TypeTalker was rated as high as that of the text editing interface (p = .16).

7.4 Discussion

Based on the results, we discuss core questions about the use of multimodal annotation for peer discussion in MOOCs, as follows: (1) what are the primary factors of modality choice, (2) what are the observed benefits of using voice over text, and (3) what can we do to convert text-major users to voice-major?

Editing and revising

The survey results suggested that the students felt that editing spoken comments was as easy as editing text. This result indicates that the TypeTalker users perceived its editing interface as effective as the conventional textbox interface, confirming the design idea that using synthesized voice will significantly reduce the user workload for editing speech. Conversely, the still low overall user satisfaction for voice, especially for production of speech comments, indicates that the difficulty in speech commenting is attributed primarily to the voice recording, not to the subsequent interactions, such as editing and revising. This interpretation paves the way for a future study that focuses on reducing user workloads at the moment of narration.

Speech recognition accuracy

Both quantitative and qualitative data suggest that the key to a better TypeTalker user experience in the field is to enhance speech recognition accuracy. The students gave quite low ratings for satisfaction with auto-transcription results ($M_{recognition\text{-}quality}$ =

2.71, SD = 1.19). From the qualitative responses of the survey, 11 voice-and-text respondents specifically mentioned transcription error as the major barrier to use speech commenting. This is half of the 22 students who tried using speech commenting at least once. Having many transcription errors in the TypeTalker interface can impose mental pressure to articulate syllables clearly to minimize their potential for mis-transcriptions. These implications can potentially explain why the students were more likely to use the text mode when they experienced low speech recognition rates.

By examining the contents of voice recordings, we could characterize the low audio quality as the major source of the transcription error. Considering that this study was conducted in a deployment setting where students can be anywhere using any device, there can be many different sources of noise (e.g., people chatting nearby, or fan noise of the laptop), and unlikely to match the lab setting for our past TypeTalker study where participants used a high-quality microphone in an acoustically insulated room. The real environment may accompany suboptimal microphone position (e.g., laptop mic) or background noise that can easily cause speech recognition errors.

Despite of all the challenges, a variety of technical remedies can enhance the speech recognition rate. For long-term use, the recognition engine can adapt to the user's speech by updating the personalized model with a new corpus. Crowdsourcing the captioning process can generate transcriptions with near-professional quality cheaply and quickly (Lasecki et al., 2012). If a noise-prone laptop microphone causes transcription error, we can instead guide the user to speak into the quality microphone of her smartphone, and then forward the audio stream to the laptop application.

English as a second language (ESL) students

The analysis on students' background data supported the implication from our previous deployment study that ESL students can face challenges for creating multimodal comments (Yoon et al., 2016). The low transcription accuracy can be especially problematic for ESL students because their accent can yield recognition errors as we observed from the analysis of the WER data. Moreover, text writing is a slow-paced activity that can feel more relaxed than live speech recording which is demanding in a second language. As a solution to this challenging problem, pipelining speech recognition with machine translation can enable ESL students to speak in their mother tongue as the native English speakers do, when high-quality automatic translation is made possible.

Missed richness on the consumption side

Confirming the results of the past consumption-side lab study for TypeTalker (Arawjo, Yoon, & Guimbretière, 2017), the students missed the richness of original voice when listening to the synthesized speech. They rated voice lower than text for the ease of understanding peers' ideas, and they could not see the benefits of using voice over typing (e.g., "the program seemed to just convert everything to text, so what was the point? [P6]" and "I could listen to comments, but I just couldn't see the point of the voice recording [P11]")). To motivate students to take advantage of the expressivity of speech, it is imperative to augment the speech re-synthesis process with acoustic richness of original speech so that the final narration can render subtle timings, inflections, and volume changes.

Type of discussion material

The lack of visuals from our task material was another potential negative factor for using voice. Use of deictic pointing synergizes the advantages of co-expressive speech by reducing communication efforts for describing diagrams and figures. Our task document was, however, primarily textual, which results in very low use of the gesture feature (only 3.5% of voice comments were gestured). The discussion for other course subjects that naturally embrace visual materials (e.g., mathematics, physics, or music) might induce more use of speech + gesture expression than text.

7.5 Summary

The previous chapter identified and addressed the usability problems that people face when creating speech comments. As a follow-up study, the suggested interface solution called TypeTalker was tested for its applicability for rich online peer discussion activity in a MOOC. In this chapter, we migrated the TypeTalker interface to RichReview.net and also implemented a scalable online infrastructure for the MOOC-scale deployment. We had 147 students conduct online peer discussions about a news article in one of the two experimental conditions: voice-and-text mode or text-only mode. The results suggested that using voice can better engage students in the discussion activity in comparison with text. However, we also observed that there were many practical barriers, such as a low speech recognition rate, low quality audio recording, and being non-native speaker, that made people shy away from using speech commenting. In the discussion, we drew implications of these findings for real-

world deployment of rich peer discussion systems, and suggested potential technical solutions to address them.

8 General discussion and future work

This chapter first provides points of discussion by revisiting the successes, limitations, and tradeoffs of the multimodal commenting system. We take a step back to ponder over the tradeoffs of our research practices, approaches, and achievements. The very basic contribution of this dissertation is to expand the field's knowledge on what multimodal commenting can offer, as well as its limits. In our evaluations, we found that while some people enjoyed the benefits of multimodal commenting, others were more sensitive to the limitations of rich media (e.g., non-native speakers felt high speech anxiety). In that regard, what follows covers the reasons behind their modality choices. In an attempt to explore a way to help people better use the speech modality, the following chapter suggests an optimal solution that balances richness and control of spoken content by leveraging advanced speech synthesis. This new perspective leads to the discussion about how new technologies, such as tablet inputs and speech recognition, served as a transition into the set of new multimodal interaction techniques. We also examine the remaining technical challenges in deploying multimodal commenting systems to real-world settings. Finally, we present projections of the generalizable and transferable aspects of our contributions to advancing multimodal annotations in future work.

8.1 Benefits and challenges of multimodal commenting

The literature in HCI suggests several benefits of multimodal commenting. Communicating via a richer medium is known to enhance expressivity (Chalfonte et al., 1991; Kraut et al., 1992), build trust (Bos, Olson, & Gergle, 2002), and establish a

better impression of the commentator (Oomen-Early et al., 2008). We deepened and expanded this knowledge by inventing new interaction techniques, and also trying them in new settings.

8.1.1 BENEFITS

The most distinctive benefit of our new rich commenting system was its efficacy in conveying complex ideas. The combination of inking, speech, and gesture is a new ensemble of modalities to simulate people's physical interaction in document explanation. Although the combination of inking and speech was explored by early multimodal annotation systems (Levine & Ehrlich, 1991; Wilcox et al., 1997), we are the first to introduce deictic gesture in document annotation. Also, we invented an easy-to-learn and lightweight interaction for capturing deictic gestures by tracking and recording the hovering tip of the stylus. In our evaluations (chapter 4.5, 5.1, and 5.1.5), people could gain the synergistic benefit of concurrent gesture and speech by connecting text to speech by pointing at the passages while speaking.

Another benefit of rich commenting is that it supports asynchronous collaboration. In contrast with synchronous communication methods, such as video chat, asynchronous commenting goes beyond temporal constraints. Asynchronicity played a pivotal role in the system's use in education because students could learn at their own pace by replaying recorded comments as needed. Such benefits enhanced the capacity of multimodal commenting to the extent that the students perceived the rich commenting system not only as a viable match to the traditional feedback method, but even as a superior alternative to face-to-face meetings (chapter 5.1.5); that is

considered the holy-grail of computer-mediated communication (Hollan & Stornetta, 1992).

8.1.2 CHALLENGES

The benefits of rich commenting come with several trade-offs. When it comes to the challenges of multimodal annotation, the focus of HCI researchers has always been the difficulties of accessing multimodal recording. The previous approaches to this problem offered better navigational cues for easier consumption. For instance, the cues in the previous studies have evolved from acoustic features (Arons, 1993; Hindus & Schmandt, 1992), ink strokes (Stifelman et al., 2001), and auto captions (Whittaker et al., 2002). Our work embraced and extended these approaches to the new application domain (e.g., feedback and discussion) by inventing new interaction techniques (e.g., TextTearing, TypeTalker) using new hardware (e.g., tablet and stylus) and technologies (e.g., speech recognition and synthesis). Moreover, our solution to this problem is especially novel in that we focused on offering more direct and fast access to these existing navigational cues. Specifically, our system embraced the fluid page layout feature to incorporate rich comments in the flow of text so that they would always be accessible for indexing. This means the students in our deployment study (chapter 5.1) could better understand instructor feedback by efficiently and effortlessly revisiting subsections of comments.

If the issue on the receiving-end of communication is access, the major concern on the producing-end is the difficulty of editing speech. The best practice for editing spoken comment has been to leverage auto caption words as proxies for editing the

corresponding audio snippets (Rubin et al., 2013; Whittaker & Amento, 2004). In this regard, we made a significant contribution by deepening our understanding about what causes user workloads in the speech editing process. We highlighted that transcription-based speech editing has too many interface modes, because the user should speak to insert new words and also fix transcription errors in a separate text mode. Identifying this problem of interface complexity led us to devise a new synthesis-based solution (chapter 6.2) that supports text-like editing of spoken content by pipelining speech recognition and synthesis modules.

The literature in psycholinguistics and communication (Brown, Fuller, & Vician, 2004; Holzman & Rousey, 1966; Nass & Brave, 2006) suggests that the difficulty of speaking is as much a psychological as a functional problem. We explored this psychological issue of speaking in the context of speech commenting. From the deployment study with peer discussion (chapter 6.1), we learned that students are concerned about having disfluencies in their recordings, and feel uncomfortable listening to their own voice. The latter point inspired us to look for a new solution that substitutes the user's voice with the machine's synthesized voice. It is worth noting here that we could address both the psychological problem and the editing problem with one solution called TypeTalker. The results of the evaluation of this solution (chapter 6.2.5) showed that using the TypeTalker interface can reduce user workload of speech editing as well as anxiety of live recording.

Reasons for modality preferences

The results of our deployment studies suggest that there are types of users who are specifically vulnerable to the challenges of multimodal annotation. For them, the cost of adapting to speech supersedes the benefits of using rich expression. Accordingly, they were more likely to prefer text commenting than multimodal commenting. A silver lining is that these groups' modality preferences were very specific to their demographic characteristics, such as English preferences or prior experience with public speech.

For instance, speaking in a foreign language is an anxiety provoking experience for most people (Horwitz, Horwitz, & Cope, 1986). The results of our deployment studies (chapter 6.1 and 7) have shown that the same holds when ESL students use speech commenting. Moreover, their issue was multilayered. First, most ESL students found it harder to speak than write in English. In addition, they were concerned about making grammatical errors, because editing speech is more effortful than editing text. Furthermore, having an accent could lower the speech recognition rate when they use our synthesis-based speech commenting interface. A potential solution is to pipeline language translation with speech recognition so that ESL students can speak in their mother tongue. This approach demands technical advancements of translation and recognition engines, because otherwise errors could accumulate as the spoken message travelled through the system.

8.2 Seeking an optimal solution: richness and control of speech commenting interfaces

While developing versions of the rich commenting system, we had the chance to explore different types of speech interfaces and examine their benefits and limitations. As shown in Figure 39, we took a fresh perspective on each type of interface to account for their trade-offs based on the framework of richness and control. Ultimately, this discussion aims to find the optimal strategy to serve the consumer and producer in speech commenting.

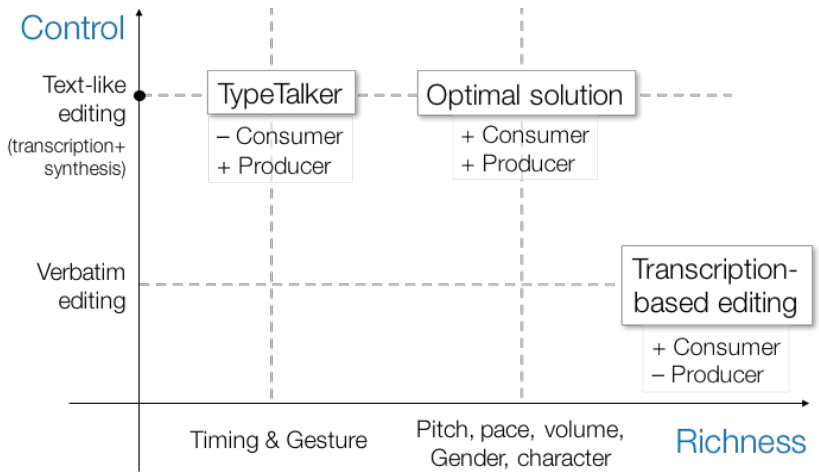


Figure 39. A design space of speech commenting user interfaces. Balancing richness and control of spoken contents is vital for both consumers and producers in communication.

The richness of speech commenting is linked to the quality of the original narration retained in the final recording. For example, a vanilla speech commenting system (e.g., voice recording feature of Adobe Reader) is on the right end of the richness spectrum because it captures all the acoustic details of speech, and simply replays it on the recipient side. The problem is that such a simple interface does not

support any control on the production side in terms of easy revision and editing of the spoken content. To enhance the control of speech commenting, Whitaker et al. (2004) and Sivaraman et al. (2016) suggest transcription-based editing interfaces (the right end of Figure 39). Using transcription enabled efficient deletion and switching of speech snippets using the captions as the semantic and visual proxy to audio stream. However, such verbatim editing could not offer quick insertion of new speech segments. Moreover, our evaluation in chapter 6.1 found that retaining minor acoustic details of the original speech caused anxiety and self-affective disturbance.

The major contribution of our TypeTalker design empowers producers of speech comments to use text-like editing. What made this possible is our novel re-synthesis approach that allows typewritten insertion of speech segments. Also, it could reduce a speaker's anxiety since the producer knows that the resulting comment will be narrated by the synthesis engine that does not have any speech disfluency. However, it will be pointless to use a speech commenting interface that simply pipelines transcription and synthesis engines (the dot on the control axis of the Figure 39), because it will lose the acoustic richness and expressivity of the original speech, which are the exact reasons people want to use speech. In TypeTalker, we transferred some richness including pause timing and gestural expression from the original narration to the synthesized one, but the results of the consumer-side lab study and the MOOC deployment study showed that the recipients still missed the full richness of the original.

Fortunately, recent advancements in speech recognition and synthesis research has opened opportunities for exploring the optimal solution that can provide both richness

and control. For example, if we exploit the latest voice conversion techniques that can profile a speaker's speech using very small corpora (Jin, Finkelstein, DiVerdi, Lu, and Mysore, 2016; Jin, Mysore, Diverdi, Lu, & Finkelstein, 2017), the synthesized narration can retain the identity of the original speaker even after editing its content. We can also retain acoustic richness (e.g., pitch, pace, volume) of original speech by taking the same approach we used for transferring the pause timing and gesture. Collectively, this optimal solution can satisfy both ends of speech commenting by giving full control of the spoken content to the commentator and also offering full richness to the recipient.

8.3 Technical challenges of deploying multimodal commenting systems in the wild

Our new interaction techniques for multimodal commenting were largely made possible by several technical advancements. Speech recognition was made cost-effective and accurate thanks to the advancements of machine learning and cloud computing. The deictic gesture feature of RichReview was made possible by integrating hovering of the tablet stylus. HTML5 enabled the development of our cross-platform web application that supports multimedia creation and transaction for a large number of users. The latest dissemination of these enabling technologies made the successful deployment of our rich feedback system possible.

However, the lack of technology can conversely work as a barrier to people's access to new systems. For instance, the last MOOC deployment study informed us that there are key technical challenges of speech commenting in real-world settings.

The low speech recognition rate could cause excessive effort to fix transcription errors, which might supersede the benefits of lightweight text-like editing. Also, even a high-quality recognizer could yield many errors when the raw audio recording has a low signal-to-noise ratio due to surrounding noise or suboptimal microphone position.

Fortunately, in response to the market's need for voice assistant and speech user interfaces, both academia and industry are rapidly advancing a set of core technologies for capturing speech: human-level speech recognition, array microphone, and noise cancelling (Ruan, Wobbrock, Liou, Ng, & Landay, 2016; Zhang & Pai, 2006). These technical advancements will lead to a much broader acceptance of rich commenting systems in the wild.

8.4 Future work

This thesis explored the topics of multimodal annotation in educational settings. Although advancing classroom technology is important research, the lessons we learned from our studies are not limited to pedagogical applications. Our aim is that this discussion creates opportunities for future work that can be extended and applied to new settings, platforms, and novel technologies.

8.4.1 IMPLICATIONS FOR EDUCATIONAL TECHNOLOGY APPLICATIONS

Our research program is unique in that we chose to deploy the system to the existing tasks in a classroom (e.g., instructor feedback and peer discussion) rather than asking people to work on a new task. This approach has a strong impact on a broad user base because we suggest a better way to accomplish the tasks that people care about. In the future, we will continue to seek opportunities to apply this rich

collaboration approach to enhance different types of educational activities, including classroom presentation, design feedback, and course evaluation. For example, end-of-term surveys are mostly in textual form because of the concern that using rich media (e.g., voice) might expose the identities of the responding students. To accommodate both richness and anonymity of survey responses, one can apply our synthesis-based speech commenting interface that enables students to expressively speak while disguising their identity with a machine-generated voice.

Giving emotional support to students has the potential to make a significant impact in the modern educational sector where many students are in emotional distress (Mahmoud, Staten, Hall, & Lennie, 2012). Several results from our study indicated the power of rich commenting to moderate and reduce negative emotions: math students felt valued as a person, because the nuances of voice empowered the instructor to moderate her negative emotions; and the TypeTalker study suggested that the carefully designed speech interface can reduce the students' anxiety and self-consciousness. The interesting research problem is that the emotional response from an instructor to a student is twofold: it can either encourage or discourage learning significantly. While our past studies explored the systems with richer media characteristics for enhanced emotional support, we want to move beyond the adjustment of media characteristics and explore active intervention techniques. In the future, we will work on building classroom communication systems that can intelligently capture and interpret on-the-fly emotional outbursts of instructors and students to give personalized feedback for their own reflection in educational activities such as feedback or discussion.

8.4.2 BEYOND DOCUMENTS

So far, the type of workspace in our thesis system was limited to documents, because classroom conversations are centered primarily around texts, such as reading materials or student submissions. We envision expanding the use cases of multimodal annotation to settings other than education. In the different use cases, collaborators might communicate over workspaces other than the documents (e.g., images for design, codes for code review, and virtual environments for 3D modelling). In that regard, supporting rich commenting on other types of workspace emerges as a major avenue of future work.

Several interactive systems have been built to offer rich commenting on types of media, including design sketches (Li, Cao, Paolantonio, & Tian, 2012), presentation slides (Kim, Glassman, Monroy-Hernández, & Morris, 2015), and virtual environment (Tsang et al., 2002). Although each previous work has investigated adapting the system to the specific target use case, their research inquiries were missing the core usability considerations for multimedia annotation, such as lightweight consumption and support for anxiety-free production, as illustrated in chapter 8.2. Hence, our future work can start from the baseline, that is to transfer the design solutions of this thesis to new contexts.

However, even after making baseline improvements, new design challenges will appear as the collaboration workspace moves to the new medium. The key to the media-specific problems lies in the attributes of the target media (e.g., visual clutter of text documents can be solved by dynamically adjusting the flow of texts). Below are

two examples of media types—code for code review and virtual environments for 3D modeling—to present potential issues and approaches.

Code review

Implementing RichReview-like multimodal commenting as a part of a software integrated development environment can enrich the code review process by allowing developers to point and speak over shared code snippets. Unlike static documents, software codes are dynamic in that they are executable and frequently changing. The problem is that a code reviewer not only talks about text of a code snippet, but also relates the code to running software (e.g., "update of this line of code caused a bug that disables this button"). Therefore, supporting rich expressions over screenshots/videos of the software as a part of multimodal commenting will be an essential way to construct a multimodal code review that connects the code and the executable.

Mixed reality applications

The devices and software for mixed reality (MR, i.e., virtual or augmented reality) are inherently designed to capture and reproduce bodily, personal, and immersive interactions. Previous studies on collaborative virtual environments focused primarily on supporting synchronous collaboration between distant workers. Here we see fresh opportunities for exploring MR-based asynchronous collaboration systems that convey the presence of co-workers across time constraints. A new challenge in consuming 3D multimodal comments is the design of navigational cues that demands a 4D representation (3D + timeline), which is difficult to comprehend. This will require

design and development of novel visualization techniques that can reduce the dimension of the cues and minimize their visual clutter in 3D. For example, one might compose and visualize a series of 3D snapshots, as with multiple exposure photography (e.g., capturing the moment when the speaker offers a deictic pronoun, or when a significant change of the scene is detected).

Such extensions of our approach for rich commenting to other media types have the potential to change how online collaborators work together in a wide variety of domains in design, business, military, and education.

9 Summaries of contributions and concluding remarks

Through the past chapters, we demonstrated an evolution of a system of work centered around the thesis that new multimodal interaction techniques will make document commenting a viable alternative to face-to-face conversation in educational settings. Throughout the iterative design process, versions of the system have been built and tested to offer the three types of generalizable and unifying contributions: design contributions for taking new approaches to usability problems, technical contributions for operationalizing the design concepts into functioning features, and empirical contributions for deriving several design implications from the lab experiments and deployment studies. This chapter chronicles the contributions of each chapter.

One foundational problem of digital annotation interfaces was that a document page often runs out of white space for commenting. TextTearing interaction incorporated visually rich annotation components (e.g., ink writing, waveform, and captions) as integral parts of a fluid page layout. We designed a variety of pen + touch gestures for quick and lightweight adjustment of the dynamic page structure. The usability testing for the different TextTearing gestures helped inform our decision to use the pen-only pigtail gesture which was the fastest and the most preferred option. To make the TextTearing interaction possible, our in-house page layout engine analyzed and restructured layouts of PDF documents. In the subsequent studies,

TextTearing served as a cornerstone interaction in the design and development of our thesis system.

To replicate rich communication modes of face-to-face meetings in mediated communication, we built a tablet-based RichReview application that exploits the full interaction capacities of speech, touch, and stylus interactions. The design of the system was geared toward solving usability issues for consuming and producing multimodal annotations. On the consumption side, a rich visual interface helped users browse audio recordings, but there was a trade-off that the visual interface overlaps the surrounding text. We solved this problem by exploiting a TextTearing interaction that incorporates the rich comments as in-line visual proxies (e.g., waveform or captions) that are free from the overlap problem while enabling the fast and direct tap-and-play style time-indexing interactions. On the production side, mixing the multiple different interaction modalities with multimedia contents might increase the user's cognitive overhead by overcomplicating the interface. To support simple and fluid interactions that reflect people's behaviors in face-to-face meetings, we made a core technical contribution that minimizes interface complexity by maintaining each visual entity (e.g., waveform, whitespace, and body text) as a first-class citizen that shares the same interface toward the modality agnostic annotation operation (e.g., speech comment, inking, and TextTearing). In the formative evaluation in a lab setting, participants found it easy and effective to convey nuanced and complex ideas using multimodal annotations.

To test the real-world efficacy of the multimodal commenting system, we recruited a small social science class where we deployed RichReview for the essay feedback

process, which is a common document-centered collaboration activity in educational settings. Since it was too costly to provide a tablet computer to each student in the class, we built a web viewer system that allowed each student to listen to the instructor's RichReview comments using their own laptop or desktop. The results showed that students preferred RichReview over traditional feedback methods, because the instructor's voice could convey rich nuances and emotions that eventually helped students grasp the full understanding of the instructor's spoken comments. Withal, some students even felt that RichReview feedback was easier to understand than face-to-face feedback, thanks to its asynchronicity and the lightweight indexing features that allowed them to review the instructor's comments at their own pace effectively.

The follow-up deployment study was targeted to extend the use of rich feedback to assignment and exam feedback in a large math class. In the week-long preliminary deployment, we learned that the major challenge for the large-scale deployment was to generate digital copies of the hardcopy submissions. For efficient digitization of the large number of math submissions, we built a semi-automatic scanning workflow where a batch process identifies page layout of the scanned images and uploads PDFs to the course management system. Students were split into two experimental groups to evaluate perceived efficacy of RichReview feedback in comparison with the traditional longhand writing on paper. From the survey results, we found that the advantages of rich feedback were best appreciated when the subject matter was complex and nuanced to the extent that describing it required lengthy recording.

In the next study, we explored the use of multimodal commenting for online peer discussion, a new pedagogy that accompanies document-centered collaboration. In contrast to the instructor feedback setting, the peer discussion setting required students not only to consume but to create rich comments. Hence, we transferred rich annotation features of the tablet app (e.g., speech recoding, and gesturing) to the web app (RichReview.net) so that students could exploit the multimodal commenting feature using microphone and mouse inputs. The peer discussion system was deployed to a small social science class for a weekly online meeting. The qualitative investigation on the students' experience informed us that speech anxiety and self-affective disturbance are the significant barriers for students to make spoken comments. Finding these new problems motivated us to take a new approach to voice production in the next study.

To solve the speech production problems, we designed, built, and evaluated TypeTalker, a novel interface for anxiety-free speech commenting. From the previous deployment, we learned that people do not enjoy listening to their own voice. This affective disturbance was the major pushback for student users to use speech for online peer discussion. This led us to invent a new approach that first transcribes a narration into text, and then lets a speech synthesis engine speak in lieu of the original speaker. This re-synthesis approach could remove self-affective disturbance by substituting the user's voice with a machine voice. Furthermore, with the new approach, editing spoken content became as easy as editing the transcription just like text, which is much less effort than the previous approach where the user had to rerecord every new snippet. To transfer rich quality of speech from the initial

recording to the synthesized speech, the time alignment module creates temporal correspondence between the speech before and after editing, and then maps pause timings and gesture segments from the original to the synthesized one, so that the synthesized narration could be a sound-alike of the speaker's voice. The evaluation showed that the new approach offers anxiety-free production of voice comments thanks to the anonymized voice and text-like editing feature. However, the users who listen to the synthesized speech missed the acoustic richness of the original narration.

Lastly, we tested the efficacy of multimodal commenting for online peer discussion by deploying a TypeTalker-laden version of RichReview to a MOOC hosted by CornellX/edX. For this MOOC-scale deployment, we implemented a scalable and secured cloud infrastructure that enables efficient transactions of multimedia data. For this week-long deployment, 110 students conducted online peer discussions using the RichReview peer discussion system. On the one hand, from users who faced significant barriers to using the multimodal commenting, we could learn that the widespread use of such a system requires accurate speech recognition, and an advanced speech synthesis engine that preserves the richness of original speech. On the other hand, when students overcome all the huddles and start to use speech, the rich commenting can motivate them to engage with the discussion activity more actively than the non-speech users.

Overall, this series of studies has deepened our understanding of why multimodal annotation is an effective way to collaborate over shared document, what are the barriers to dissemination of the rich commenting, and how to tackle such challenges with new design approaches and technical solutions. Then we critically evaluated the

abovementioned contributions to examine tradeoffs of our solutions and to open opportunities for future work. We envisioned the optimal speech commenting interface that balances the benefit for commentators and listeners. To support users working in a silent scholarly environment, we suggested quiet multimodal annotations, a new genre of rich commenting that leverages speechless communication modalities to augment textual annotation. In the long term, our approaches to multimodal annotation can be extended to new computational platforms, such as virtual or augmented reality, or new applications for advancing technologies in the classroom.

REFERENCES

Ades, S., & Swinehart, D. C. (1986). *Voice annotation and editing in a workstation environment*. XEROX Corporation, Palo Alto Research Center.

Adobe. (2016a). Acrobat Reader. Retrieved October 9, 2016, from https://get.adobe.com/reader/

Adobe. (2016b). Audition.

Adobe. (2016c). Premiere. Retrieved May 18, 2016, from http://www.adobe.com/products/premiere.html

Agrawala, M., & Shilman, M. (2005). DIZI: a digital ink zooming interface for document annotation. *Human-Computer Interaction-INTERACT 2005*. Retrieved from http://link.springer.com/chapter/10.1007/11555261_9

Anderson, R. (2004). Beyond PowerPoint: Building a new classroom presenter. *SYLLABUS-SUNNYVALE THEN* …. Retrieved from http://dada.cs.washington.edu/dl/presenter/papers/2004/SYLLABUS_2004.pdf

Anderson, R., Chung, O., Davis, K. M., Davis, P., Prince, C., Razmov, V., & Simon, B. (2006). Classroom Presenter-A Classroom Interaction System for Active and Collaborative Learning. *Proceedings of the 1st Workshop on the Impact of Pen-Based Technology on Education (WIPTE06)*. Retrieved from http://www.craigprince.com/papers/AACDDPRS_WIPTE_2006.pdf

Anson, C. M. (1997). In our own voices: Using recorded commentary to respond to writing. *New Directions for Teaching & Learning*, *69*(Spring), 105–113. https://doi.org/10.1002/tl.6909

Apple. (2016). Garage Band. Retrieved May 18, 2016, from http://www.apple.com/ios/garageband

Arawjo, I., Yoon, D., & Guimbretière, F. (2017). TypeTalker: A Speech Synthesis-Based Multi-Modal Commenting System. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work & Social Computing*.

Arons, B. (1993). SpeechSkimmer: Interactively Skimming Recorded Speech. In *Proceedings of the 6th Annual ACM Symposium on User Interface Software and Technology* (pp. 187–196). New York, NY, USA: ACM. https://doi.org/10.1145/168642.168661

Audacity Team. (2016). Audacity. Retrieved May 18, 2016, from http://www.audacityteam.org

Barber, M., Donnelly, K., Rizvi, S., & Summers, L. (2013). An avalanche is coming: higher education and the revolution ahead. *The Institute for Public Policy Research*. Retrieved from http://www.voced.edu.au/content/ngv55590

Bekker, M. M., Olson, J. S., & Olson, G. M. (1995). Analysis of gestures in face-to-face design teams provides guidance for how to use groupware in design. In *Proceedings of the conference on Designing interactive systems processes, practices, methods, & techniques - DIS '95* (pp. 157–166). New York, New

York, USA: ACM Press. https://doi.org/10.1145/225434.225452

Berthouzoz, F., Li, W., & Agrawala, M. (2012). Tools for placing cuts and transitions in interview video. *ACM Transactions on Graphics*, *31*(4), 1–8. https://doi.org/10.1145/2185520.2185563

Bickmore, T., Pfeifer, L., & Yin, L. (2008). the Role of Gesture in Document Explanation By Embodied Conversational Agents. *International Journal of Semantic Computing*, *2*(1), 47–70. https://doi.org/10.1142/S1793351X08000348

Birnholtz, J., Steinhardt, S., & Pavese, A. (2013). Write here, write now! In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (p. 961). New York, New York, USA: ACM Press. https://doi.org/10.1145/2470654.2466123

Blurton-Jones, N. G. (1972). Non-verbal communication in children. *Nonverbal Communication*, 271–296.

Bos, N., Olson, J., & Gergle, D. (2002). Effects of four computer-mediated communications channels on trust development. *Proceedings of the SIGCHI …*, *4*(July 2015), 135–140. https://doi.org/10.1145/503376.503401

Brown, S. A., Fuller, R. M., & Vician, C. (2004). WHO'S AFRAID OF THE VIRTUAL WORLD? Anxiety and Computer-Mediated Communication. *Journal of the Association for Information Systems*, *5*(2), 79–107. Retrieved from http://ezproxy.lib.ucf.edu/login?URL=http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=16585708&site=ehost-live

Brush, A., Bargeron, D., & Grudin, J. (2002). Supporting interaction outside of class: anchored discussions vs. discussion boards. *Computer Cupport for Collaborative Learning: Foundation for a CSCL Communiy*, 425–434. Retrieved from http://dl.acm.org/citation.cfm?id=1658676

Cambridge, U. of. (2016). HTK. Retrieved May 26, 2016, from http://htk.eng.cam.ac.uk/

Chalfonte, B. L., Fish, R. S., & Kraut, R. E. (1991). Expressive richness: A COMPARISON OF SPEECH AND TEXT AS MEDIA FOR REVISION. In *Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91* (pp. 21–26). New York, New York, USA: ACM Press. https://doi.org/10.1145/108844.108848

Chang, B.-W., Mackinlay, J. D., Zellweger, P. T., & Igarashi, T. (1998). A negotiation architecture for fluid documents. In *Proceedings of the 11th annual ACM symposium on User interface software and technology - UIST '98* (pp. 123–132). New York, New York, USA: ACM Press. https://doi.org/10.1145/288392.288585

Chen, N., Guimbretiere, F., & Sellen, A. (2012). Designing a multi-slate reading environment to support active reading activities. *ACM Transactions on Computer-Human Interaction*, *19*(3), 1–35. https://doi.org/10.1145/2362364.2362366

Clark, H. H. (1996). *Using Language*. *Computational Linguistics* (Vol. 23). https://doi.org/10.2277/0521561582

Clow, D. (2013). MOOCs and the Funnel of Participation. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 185–189). New York, NY, USA: ACM. https://doi.org/10.1145/2460296.2460332

Cockburn, A., Gutwin, C., & Alexander, J. (2006). Faster document navigation with space-filling thumbnails. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '06*. Retrieved from http://dl.acm.org/citation.cfm?id=1124774

Coetzee, D., Fox, A., Hearst, M. A., & Hartmann, B. (2014). Should Your MOOC Forum Use a Reputation System? In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1176–1187). New York, NY, USA: ACM. https://doi.org/10.1145/2531602.2531657

Coetzee, D., Lim, S., Fox, A., Hartmann, B., & Hearst, M. A. (2015). Structuring Interactions for Large-Scale Synchronous Peer Learning. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1139–1152). New York, NY, USA: ACM. https://doi.org/10.1145/2675133.2675251

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition*, *106*(2), 1047–58. https://doi.org/10.1016/j.cognition.2007.04.010

CornellX. (2016). GMO101x The Science and Politics of the GMO. Retrieved June 4, 2017, from https://courses.edx.org/courses/course-v1:CornellX+GMO101x+3T2016/info

Cox, B., & Cox, B. (2008). Developing interpersonal and group dynamics through asynchronous threaded discussions: The use of discussion board in collaborative learning. *Education*, *128*, 553–565.

Creswell, B. J. W., Piano, V. L., & Published, C. (2007). Designing and Conducting Mixed Methods Research. *Australian and New Zealand Journal of Public Health*, *31*(4), 388–388. https://doi.org/10.1111/j.1753-6405.2007.00096.x

Crook, A., Mauchline, A., Maw, S., Lawson, C., Drinkwater, R., Lundqvist, K., … Park, J. (2012). The use of video technology for providing feedback to students: Can it enhance the feedback experience for staff and students? *Computers and Education*, *58*(1), 386–396. https://doi.org/10.1016/j.compedu.2011.08.025

Daft, R. L., & Lengel, R. H. (1986). Organizational Information Requirements, Media Richness and Structural Design. *Management Science*, *32*(5), 554–571. https://doi.org/10.1287/mnsc.32.5.554

Duda, R. O., & Hart, P. E. (1972). Use of the Hough Transformation to Detect Lines and Curves in Pictures. *Commun. ACM*, *15*(1), 11–15. https://doi.org/10.1145/361237.361242

Fischler, M. A., & Bolles, R. C. (1981). Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM*, *24*(6), 381–395. https://doi.org/10.1145/358669.358692

Fussell, S. R., Setlock, L. D., Yang, J., Ou, J., Mauer, E., & Kramer, A. D. I. (2004). Gestures over Video Streams to Support Remote Collaboration on Physical

Tasks. *Hum.-Comput. Interact.*, *19*(3), 273–309. https://doi.org/10.1207/s15327051hci1903_3

Glaser, B. G., & Strauss, A. L. (1967). The discovery of grounded theory. *International Journal of Qualitative Methods*, *5*, 1–10. https://doi.org/10.2307/588533

Goldin-Meadow, S. (2005). *Hearing gesture: How our hands help us think*. Harvard University Press.

Golovchinsky, G., & Denoue, L. (2002). Moving markup: Repositioning Freeform Annotations. In *Proceedings of the 15th annual ACM symposium on User interface software and technology - UIST '02* (p. 21). New York, New York, USA: ACM Press. https://doi.org/10.1145/571985.571989

Golovchinsky, G., Price, M. N., & Schilit, B. N. (1999). From reading to retrieval: freeform ink annotations as queries. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99* (pp. 19–25). New York, New York, USA: ACM Press. https://doi.org/10.1145/312624.312637

Grudin, J. (1988). Why CSCW applications fail: problems in the design and evaluationof organizational interfaces. In *Proceedings of the 1988 ACM conference on Computer-supported cooperative work* (pp. 85–93). New York, New York, USA: ACM Press. https://doi.org/10.1145/62266.62273

Guiard, Y. (1987). Asymmetric division of labor in human skilled bimanual action: the

kinematic chain as a model. *Journal of Motor Behavior*, *19*(4), 486–517. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15136274

Gutwin, C., Stark, G., & Greenberg, S. (1995). Support for Workspace Awareness in Educational Groupware. In *Proc Conference on Computer Spported Collaborative Learning* (pp. 147–156). Morristown, NJ, USA: Association for Computational Linguistics. https://doi.org/10.3115/222020.222126

Haberman, C. (2014). The Head-Scratching Case of the Vanishing Bees. Retrieved June 4, 2017, from https://www.nytimes.com/2014/09/29/us/the-head-scratching-case-of-the-vanishing-bees.html?_r=0

Hardock, G., Kurtenbach, G., & Buxton, W. (1993). A marking based interface for collaborative writing. In *Proceedings of the 6th annual ACM symposium on User interface software and technology - UIST '93* (pp. 259–266). New York, New York, USA: ACM Press. https://doi.org/10.1145/168642.168669

Harrison, S., Minneman, S., & Marinacci, J. (1999). The DrawStream Station or the AVCs of Video Cocktail Napkins. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems - Volume 2* (pp. 543–549). Washington, DC, USA: IEEE Computer Society. https://doi.org/10.1109/MMCS.1999.779259

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, *52*(C), 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hew, K. F., & Cheung, W. S. (2012). Examining the use of asynchronous voice discussion in a blended-learning environment. *Proceedings of the European Conference on E-Government, ECEG*, *10*(4), 136–140.

Hew, K. F., & Cheung, W. S. (2013). Audio-based versus text-based asynchronous online discussion: Two case studies. *Instructional Science*, *41*(2), 365–380. https://doi.org/10.1007/s11251-012-9232-7

Hinckley, K., Baudisch, P., Ramos, G., & Guimbretiere, F. (2005). Design and analysis of delimiters for selection-action pen gesture phrases in scriboli. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '05* (p. 451). New York, New York, USA: ACM Press. https://doi.org/10.1145/1054972.1055035

Hinckley, K., Bi, X., Pahud, M., & Buxton, B. (2012). Informal information gathering techniques for active reading. *Proceedings of the 2012 ACM …*. Retrieved from http://dl.acm.org/citation.cfm?id=2208327

Hinckley, K., Yatani, K., Pahud, M., Coddington, N., Rodenhouse, J., Wilson, A., … Buxton, B. (2010). Pen + touch = new tools. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology - UIST '10* (p. 27). New York, New York, USA: ACM Press. https://doi.org/10.1145/1866029.1866036

Hinckley, K., Zhao, S., Sarin, R., Baudisch, P., Cutrell, E., Shilman, M., & Tan, D. (2007). InkSeine. In *Proceedings of the SIGCHI conference on Human factors in*

computing systems - CHI '07 (p. 251). New York, New York, USA: ACM Press. https://doi.org/10.1145/1240624.1240666

Hindus, D., & Schmandt, C. (1992). Ubiquitous audio: Capturing Spontaneous Collaboration. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work - CSCW '92* (pp. 210–217). New York, New York, USA: ACM Press. https://doi.org/10.1145/143457.143481

Hollan, J., & Stornetta, S. (1992). Beyond Being There. In *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems* (pp. 119–125). New York, NY, USA: ACM. https://doi.org/10.1145/142750.142769

Holzman, P. S., & Rousey, C. (1966). The voice as a percept. *Journal of Personality and Social Psychology*, *4*(1), 79–86. https://doi.org/10.1037/h0023518

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). *Foreign Language Classroom Anxiety*. *The Modern Language Journal* (Vol. 70). https://doi.org/10.1111/j.1540-4781.1986.tb05256.x

Huang, J., Dasgupta, A., Ghosh, A., Manning, J., & Sanders, M. (2014). Superposter behavior in MOOC forums. *Proceedings of the First ACM Conference on Learning @ Scale Conference - L@S '14*, 117–126. https://doi.org/10.1145/2556325.2566249

IBM, I. (n.d.). IBM Bluemix. Retrieved May 26, 2016, from http://www.ibm.com/Bluemix

Ice, P., Curtis, R., Phillips, P., & Wells, J. (2007). Using Asynchronous Audio Feedback to Enhance Teaching Presence and Students' Sense of Community. *Journal of Asynchronous Learning Networks*, *11*(2), 3–25.

Jin, Z., Finkelstein, A., DiVerdi, S., Lu, J., and Mysore, G. J. (2016). CUTE: a concatenative method for voice conversion using exemplar-based unit selection. In *In 41st IEEE International Conference on Acoustics Speech and Signal Processing*. IEEE.

Jin, Q., Toth, A. R., Schultz, T., & Black, A. W. (2009). Speaker de-identification via voice transformation. In *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009* (pp. 529–533). https://doi.org/10.1109/ASRU.2009.5373356

Jin, Z., Mysore, G. J., Diverdi, S., Lu, J., & Finkelstein, A. (2017). VoCo: Text-based Insertion and Replacement in Audio Narration.

Karam, M., & Schraefel, m. c. (2005). *A Taxonomy of Gestures in Human Computer Interactions*. *Technical Report, Eletronics and Computer Science*. https://doi.org/10.1.1.97.5474

Kim, J., Glassman, E. L., Monroy-Hernández, A., & Morris, M. R. (2015). RIMES: Embedding Interactive Multimedia Exercises in Lecture Videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15* (pp. 1535–1544). New York, NY, USA: ACM. https://doi.org/10.1145/2702123.2702186

Kita, S. (2003). Pointing: A foundational building block of human communication. In *Pointing: Where language, culture, and cognition meet* (pp. 1–8). https://doi.org/10.4324/9781410607744

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 170–179). New York, NY, USA: ACM. https://doi.org/10.1145/2460296.2460330

Klammer, E. (1973). Cassettes in the Classroom. *College English*, *35*(2), 179–189. Retrieved from http://www.jstor.org/stable/375445

Kraut, R., Galegher, J., Fish, R., & Chalfonte, B. (1992). Task Requirements and Media Choice in Collaborative Writing. *Human-Computer Interaction*, *7*(4), 375–407. https://doi.org/10.1207/s15327051hci0704_2

Kulkarni, C., Cambre, J., Kotturi, Y., Bernstein, M. S., & Klemmer, S. R. (2015). Talkabout: Making Distance Matter with Small Groups in Massive Classes. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15* (pp. 1116–1128). New York, New York, USA: ACM Press. https://doi.org/10.1145/2675133.2675166

Kurtenbach, G. P. (1993). *The design and evaluation of marking menus*. University of Toronto.

Kuzuoka, H. (1992). Spatial workspace collaboration: a SharedView video support

system for remote collaboration capability. *Proceedings of the ACM Conference on Human Factors in Computing Systems*, *Monterey,* 533–540. https://doi.org/10.1145/142750.142980

Lasecki, W., Miller, C., Sadilek, A., Abumoussa, A., Borrello, D., Kushalnagar, R., & Bigham, J. (2012). Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology - UIST '12* (p. 23). New York, NY, USA: ACM. https://doi.org/10.1145/2380116.2380122

Lee, J.-S., & Tatar, D. (2012). "Good enough" pointing in pervasive computing. In *Collaboration Technologies and Systems (CTS), 2012 International Conference on* (pp. 330–337). https://doi.org/10.1109/CTS.2012.6261071

Levine, S. R., & Ehrlich, S. F. (1991). The Freestyle System. In *Human-Machine Interactive Systems* (pp. 3–21). Springer.

Li, G., Cao, X., Paolantonio, S., & Tian, F. (2012). SketchComm. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12* (p. 359). New York, New York, USA: ACM Press. https://doi.org/10.1145/2145204.2145261

Lofland, J., & Lofland, L. (1971). Analyzing social settings. *Belmont, CA: Wadsworth*.

Mahmoud, J., Staten, R., Hall, L., & Lennie, T. (2012). The Relationship among Young Adult College Students' Depression, Anxiety, Stress, Demographics, Life Satisfaction, and Coping Styles. *Issues in Mental Health Nursing*, *33*(2006), 149–

156. https://doi.org/10.3109/01612840.2011.632708

Mak, S. F. J., Williams, R., & Mackness, J. (2010). Blogs and forums as communication and learning tools in a MOOC. *Proceedings of the 7th International Conference on Networked Learning*, 275–285. https://doi.org/10.1162/dmal.9780262524834.071

Marriott, P., & Hiscock, J. (2002). Voice vs text-based discussion forums: An implementation of Wimba Voice Boards. In *Proceedings of world conference on e-learning in corporate, government, healthcare, and higher education 2002* (Vol. 2002, pp. 640–646).

Marshall, C. C. (1997). Annotation: from paper books to the digital library. In *Proceedings of the second ACM international conference on Digital libraries - DL '97* (pp. 131–140). New York, New York, USA: ACM Press. https://doi.org/10.1145/263690.263806

Marshall, C. C. (1998). Toward an ecology of hypertext annotation. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space---structure in hypermedia systems links, objects, time and space---structure in hypermedia systems - HYPERTEXT '98* (pp. 40–49). New York, New York, USA: ACM Press. https://doi.org/10.1145/276627.276632

Marshall, C. C., & Brush, A. J. B. (2004). Exploring the relationship between personal and public annotations. *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, 2004*. https://doi.org/10.1109/JCDL.2004.1336148

Marshall, C. C., Price, M. N., Golovchinsky, G., & Schilit, B. N. (1999). Collaborating over portable reading appliances. *Personal Technologies*, *3*(1–2), 43–53. https://doi.org/10.1007/BF01305319

Maxwell, J. A. (2013). Qualitative research design: An interactive approach. In *Qualitative research design: An interactive approach* (pp. 23–38). https://doi.org/10.1007/s13398-014-0173-7.2

Mayer, R. E. (2005). *The Cambridge handbook of multimedia learning*. *Cambridge handbook of multimedia learning*. https://doi.org/10.1075/idj.16.1.13pel

McNeill, D. (1992). Hand and Mind: What Gestures Reveal About Thought. *What Gestures Reveal about*, 1–15. https://doi.org/10.2307/1576015

McNeill, D. (2005). Gesture and Thought. *Continuum*, (18), University of Chicago Press. https://doi.org/10.7208/chicago/9780226514642.001.0001

Monserrat, T.-J. K. P., Zhao, S., McGee, K., & Pandey, A. V. (2013). NoteVideo: Facilitating Navigation of Blackboard-style Lecture Videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (p. 1139). New York, New York, USA: ACM Press. https://doi.org/10.1145/2470654.2466147

Morris;, M. R., Brush;, A. J. B., & Meyers, B. R. (2007). Reading revisited: Evaluating the usability of digital display surfaces for active reading tasks. *Horizontal Interactive Human-Computer System*, 79–86. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4384115

Munteanu, C., Baecker, R., Penn, G., Toms, E., & James, D. (2006). The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. *Proc. CHI '06, ACM Press*, 493. https://doi.org/10.1145/1124772.1124848

Nass, C., & Brave, S. (2006). Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship. *Computational Linguistics*. https://doi.org/10.1162/coli.2006.32.3.451

Neuwirth, C. M., Chandhok, R., Charney, D., Wojahn, P., & Kim, L. (1994). Distributed collaborative writing. In *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94* (pp. 51–57). New York, New York, USA: ACM Press. https://doi.org/10.1145/191666.191693

Ong, W. (1989). *Orality and literacy. Language & Communication* (Vol. 9). https://doi.org/10.1016/0271-5309(89)90011-6

Oomen-Early, J., Bold, M., Wiginton, K. L., Gallien, T. L., & Anderson, N. (2008). Using asynchronous audio communication (AAC) in the online classroom: A comparative study. *Journal of Online Learning and Teaching*, *4*(3), 267–276.

Pan, Y., Jiang, D., Picheny, M., & Qin, Y. (2009). Effects of real-time transcription on non-native speaker's comprehension in computer-mediated communications. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI '09* (p. 2353). https://doi.org/10.1145/1518701.1519061

Pearson, J., Buchanan, G., & Thimbleby, H. (2009). Improving annotations in digital documents. … *Advanced Technology for Digital …*. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-04346-8_51

Pearson, J., Buchanan, G., & Thimbleby, H. (2011). The reading desk: applying physical interactions to digital documents. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11* (p. 3199). New York, New York, USA: ACM Press. https://doi.org/10.1145/1978942.1979416

Price, M. N., Schilit, B. N., & Golovchinsky, G. (1998). XLibris. In *CHI 98 conference summary on Human factors in computing systems - CHI '98* (pp. 22–23). New York, New York, USA: ACM Press. https://doi.org/10.1145/286498.286510

Randles, B., Yoon, D., Cheatle, A., Jung, M., & Guimbretiere, F. (2015). Supporting Face-to-Face Like Communication Modalities for Asynchronous Assignment Feedback in Math Education. In *L@S 2015 - 2nd ACM Conference on Learning at Scale* (pp. 321–326). New York, NY, USA: ACM. https://doi.org/10.1145/2724660.2728684

Raskin, J. (2000). *The Humane Interface: New Directions for Designing Interactive Systems*. *Addison Wesley Pub Co Inc*. https://doi.org/10.1076/ilee.10.3.299.8765

Ruan, S., Wobbrock, J. O., Liou, K., Ng, A., & Landay, J. (2016). Speech Is 3x Faster than Typing for English and Mandarin Text Entry on Mobile Devices. In *arXiv preprint arXiv:1608.07323*.

Rubin, S., Berthouzoz, F., Mysore, G. J., Li, W., & Agrawala, M. (2013). Content-based tools for editing audio stories. In *Proceedings of the 26th annual ACM symposium on User interface software and technology - UIST '13* (pp. 113–122). New York, New York, USA: ACM Press. https://doi.org/10.1145/2501988.2501993

Scheier, M. F., & Carver, C. S. (1985). The Self-Consciousness Scale: A Revised Version for Use with General Populations. *Journal of Applied Social Psychology*. https://doi.org/10.1111/j.1559-1816.1985.tb02268.x

Schilit, B. N., Golovchinsky, G., & Price, M. N. (1998). Beyond paper: supporting active reading with free form digital ink annotations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 249–256). New York, NY, USA: ACM Press/Addison-Wesley Publishing Co. https://doi.org/10.1145/274644.274680

Schmandt, C. (1981). The intelligent ear: A graphical interface to digital audio. … , *IEEE International Conference on Cybernetics and …*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.147.7161

Scholl, J., McCarthy, J., & Harr, R. (2006). A comparison of chat and audio in media rich environments. *Proceedings of ACM CSCW'06 Conference on Computer-Supported Cooperative Work*, 323–332. https://doi.org/10.1145/1180875.1180925

Schröder, M. (2001). Emotional Speech Synthesis: A Review. In *INTERSPEECH* (pp.

561–564). Retrieved from http://www.isca-speech.org/archive/eurospeech_2001/e01_0561.html

Selingo, J. J. (2013). *College (un) bound: The future of higher education and what it means for students*. Houghton Mifflin Harcourt.

Sellen, A. J., & Harper, R. H. (2003). *The Myth of the Paperless Office*. *MIT Press*.

Silva, M. L. (2012). Camtasia in the Classroom: Student Attitudes and Preferences for Video Commentary or Microsoft Word Comments During the Revision Process. *Computers and Composition*, *29*(1), 1–22. https://doi.org/http://dx.doi.org/10.1016/j.compcom.2011.12.001

Sivaraman, V., Yoon, D., & Mitros, P. (2016). Simplified Audio Production in Asynchronous Voice-Based Discussions. In *Proceedings of the 2016 ACM Conference on Human Factors in Computing Systems*.

Soltau, H., Saon, G., & Sainath, T. N. (2014). Joint training of convolutional and non-convolutional neural networks. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (pp. 5572–5576). https://doi.org/10.1109/ICASSP.2014.6854669

Song, H., Benko, H., Guimbretiere, F., Izadi, S., Cao, X., & Hinckley, K. (2011). Grips and gestures on a multi-touch pen. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11* (p. 1323). New York, New York, USA: ACM Press. https://doi.org/10.1145/1978942.1979138

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: An historical perspective. *Cambridge Handbook of the Learning Sciences*, *2006*, 409–426. https://doi.org/10.1145/1124772.1124855

Stark, L., Whittaker, S., & Hirschberg, J. (2000). ASR satisficing: the effects of ASR accuracy on speech retrieval. *INTERSPEECH*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.9637&rep=rep1&type=pdf

Stifelman, L., Arons, B., & Schmandt, C. (2001). The Audio Notebook: Paper and Pen Interaction with Structured Speech. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 182–189). New York, NY, USA: ACM Press. https://doi.org/10.1145/365024.365096

Tang, J. C. (1991). Findings from observational studies of collaborative work. *International Journal of Man-Machine Studies*, *34*(2), 143–160. https://doi.org/10.1016/0020-7373(91)90039-A

Tashman, C. S., & Edwards, W. K. (2011). LiquidText: a flexible, multitouch environment to support active reading. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11* (p. 3285). New York, New York, USA: ACM Press. https://doi.org/10.1145/1978942.1979430

Tsang, M., Fitzmaurice, G. W., Kurtenbach, G., Khan, A., & Buxton, B. (2002). Boom Chameleon: Simultaneous Capture of 3D Viewpoint, Voice and Gesture Annotations on a Spatially-aware Display. In *Proceedings of the 15th Annual*

ACM Symposium on User Interface Software and Technology (pp. 111–120). New York, NY, USA: ACM. https://doi.org/10.1145/571985.572001

Tu, C.-H., & McIsaac, M. (2002). The Relationship of Social Presence and Interaction in Online Classes. *American Journal of Distance Education*, *16*(3), 131–150. https://doi.org/10.1207/S15389286AJDE1603_2

Valbret, H., Moulines, E., & Tubach, J. P. (1992). Voice transformation using PSOLA technique. *Speech Communication*, *11*(2–3), 175–187. https://doi.org/10.1016/0167-6393(92)90012-V

Vemuri, S., DeCamp, P., Bender, W., & Schmandt, C. (2004). Improving speech playback using time-compression and speech recognition. In *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04* (pp. 295–302). New York, New York, USA: ACM Press. https://doi.org/10.1145/985692.985730

Vogel, D., & Balakrishnan, R. (2010). Occlusion-aware interfaces. In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10* (p. 263). New York, New York, USA: ACM Press. https://doi.org/10.1145/1753326.1753365

Wang, Q., & Nass, C. (2005). Less Visible and Wireless: Two Experiments on the Effects of Microphone Type on Users' Performance and Perception. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 809–818). New York, NY, USA: ACM.

https://doi.org/10.1145/1054972.1055086

Whittaker, S., & Amento, B. (2004). Semantic speech editing. In *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04* (pp. 527–534). New York, New York, USA: ACM Press. https://doi.org/10.1145/985692.985759

Whittaker, S., Geelhoed, E., & Robinson, E. (1993). Shared workspaces: how do they work and when are they useful? *International Journal of Man-Machine Studies*, *39*(5), 813–842. https://doi.org/10.1006/imms.1993.1085

Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., … Rosenberg, A. (2002). SCANMail: a voicemail interface that makes speech browsable, readable and searchable. In *Proceedings of the SIGCHI conference on Human factors in computing systems Changing our world, changing ourselves - CHI '02* (p. 275). New York, New York, USA: ACM Press. https://doi.org/10.1145/503376.503426

Whittaker, S., Hyland, P., & Wiley, M. (1994). Filochat: handwritten notes provide access to recorded conversations. In *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94* (pp. 271–277). New York, New York, USA: ACM Press. https://doi.org/10.1145/191666.191763

Wilcox, L. D., Schilit, B. N., & Sawhney, N. (1997). Dynomite: a dynamically organized ink and audio notebook. In *Proceedings of the SIGCHI conference on*

Human factors in computing systems - CHI '97 (pp. 186–193). New York, New York, USA: ACM Press. https://doi.org/10.1145/258549.258700

Williams, E. (1977). Experimental Comparisons of Face-to-Face and Mediated Communication: A Review. *Psychological Bulletin*, *84*(5), 963–976. https://doi.org/10.1037/0033-2909.84.5.963

Yaneske, E., & Oates, B. (2010). Using voice boards: Pedagogical design, technological implementation, evaluation and reflections. *Australasian Journal of Educational Technology*, *26*(8), 233–250. https://doi.org/10.3402/rlt.v18i3.10767

Yoon, D., Chen, N., & Guimbretière, F. (2013). TextTearing: Opening White Space for Digital Ink Annotation. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (pp. 107–112). New York, NY, USA: ACM. https://doi.org/10.1145/2501988.2502036

Yoon, D., Chen, N., Guimbretière, F., & Sellen, A. (2014). RichReview: Blending Ink, Speech, and Gesture to Support Collaborative Document Review. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (pp. 481–490). New York, NY, USA: ACM. https://doi.org/10.1145/2642918.2647390

Yoon, D., Chen, N., Randles, B., Cheatle, A., Löckenhoff, C. E., Jackson, S. J., … Guimbretière, F. (2016). RichReview++: Deployment of a Collaborative Multi-modal Annotation System for Instructor Feedback and Peer Discussion. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative*

Work & Social Computing (pp. 195–205). https://doi.org/10.1145/2818048.2819951

Zeleznik, R., Bragdon, A., Adeputra, F., & Ko, H.-S. (2010). Hands-on math: : a page-based multi-touch and pen desktop for technical work and problem solving. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology - UIST '10* (p. 17). New York, New York, USA: ACM Press. https://doi.org/10.1145/1866029.1866035

Zhang, M., & Pai, W. C. (2006). Small array microphone for acoustic echo cancellation and noise suppression. Google Patents. Retrieved from https://www.google.com/patents/US7003099

Zheng, Q., Booth, K., & McGrenere, J. (2006). Co-authoring with structured annotations. In *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06* (p. 131). New York, New York, USA: ACM Press. https://doi.org/10.1145/1124772.1124794

Zheng, S., Rosson, M. B., Shih, P. C., & Carroll, J. M. (2015). Understanding Student Motivation, Behaviors and Perceptions in MOOCs. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1882–1895). New York, NY, USA: ACM. https://doi.org/10.1145/2675133.2675217

Zyto, S., Karger, D., Ackerman, M., & Mahajan, S. (2012). Successful classroom deployment of a social document annotation system. In *Proceedings of the*

SIGCHI Conference on Human Factors in Computing Systems (pp. 1883–1892). New York, New York, USA: ACM Press. https://doi.org/10.1145/2207676.2208326